Unit 9

Testing hypotheses

Introduction

A **hypothesis** is a statement about something that may or may not be true. Examples of such statements which have made headline news include:

- Over the festive period, the average person could gain around 2.3 kg (5 lb) in weight (claim by the British Dietetic Association reported in *The Guardian*, 20 December 2013).
- Harder driving theory test leads to falling pass rate (*BBC News*, 11 September 2014).
- Blueberries lower blood pressure in menopausal women (*New York Times*, 14 January 2015).
- One in four UK young adults aged between 20 and 34 years still live with their parents (*The Guardian*, 21 January 2014).

This unit considers how we can use data to test the validity of such hypotheses.

The first thing that a statistician needs to do when testing hypotheses is to reinterpret the hypotheses in terms of values of unknown population parameters such as the mean, μ. For example, consider the statement 'over the festive period, the average person could gain around 2.3 kg in weight'. This can be reinterpreted in terms of a hypothesis that parameter μ, the mean weight gain over the festive period, is around 2.3 kg. Section 1 looks in detail at how to specify hypotheses suitable for statistical testing.

The next step is to take a sample of data relevant to the hypotheses to be tested. So, for example, when testing the hypothesis concerning weight gain over the festive period, data need to be collected on the weight gains over the festive period of a sample from the population. Testing a hypothesis then involves investigating how likely it is that the data we actually observe would have arisen if our hypothesis is true. The main ideas of testing hypotheses are presented in Section 2, and some common tests are then presented in Section 3. The link between hypothesis testing and the confidence intervals that you studied in Unit 8 is also addressed (in Subsection 3.3).

Section 4 considers an alternative method of testing hypotheses which is based on what are known as significance probabilities, or p-values, and makes the link between the two approaches. Finally, Section 5 considers a measure of how well a hypothesis test performs, known as the power of the test, and shows how the power can guide statisticians when choosing the data sample size for a test.

1 Specifying hypotheses

A hypothesis test concerns the (unknown) value of a population parameter and usually involves two hypotheses: the null hypothesis and the alternative hypothesis.

The **null hypothesis**, denoted H_0, represents what we assume about the parameter before we observe any data. This is framed in a 'nothing is happening' kind of way even if – for example – the study is being undertaken to attempt to verify that a treatment works: the 'null' essentially stands for 'no difference'. The null hypothesis is, therefore, what we are usually attempting to find evidence against. For example, in testing the claim that blueberries lower blood pressure in menopausal women, the null hypothesis would be that blueberries make no difference to blood pressure in menopausal women. As another example, in testing the claim that the driving theory test pass rate is falling, the null hypothesis would be that the test pass rate is no different from some representative measure of previous years' pass rates, such as the pass rate in the preceding year.

The **alternative hypothesis**, denoted H_1, is simply the hypothesis that is considered as an alternative to the null hypothesis. The choice of alternative hypothesis depends on the context and purpose of the statistical test. For example, in testing the claim that blueberries lower blood pressure in menopausal women, the alternative hypothesis would be that blueberries *do* lower blood pressure in menopausal women, and when testing the claim that the driving theory test pass rate is falling, the alternative hypothesis would be that the pass rate *is* falling.

In M248, when testing hypotheses involving a single parameter, the null hypothesis will always specify a single value θ_0, say, for the unknown parameter θ. The null hypothesis is then written as

$$H_0 : \theta = \theta_0.$$

How to specify the null hypothesis for single parameter problems is illustrated in the following examples.

Example 1 *Null hypothesis: festive weight gain*

Consider once again the hypothesis that 'over the festive period, the average person could gain around 2.3 kg in weight'. As mentioned in the Introduction, in order to carry out a statistical test, this hypothesis needs to be reinterpreted in terms of unknown parameters, and the parameter μ, the mean weight gain over the festive period, was suggested.

We therefore wish to test whether μ could be around 2.3 kg. The null hypothesis representing 'no difference' suggests specifying μ_0 to be 2.3 kg so that the null hypothesis is

$$H_0 : \mu = 2.3 \text{ kg}.$$

Example 2 *Null hypothesis: proportion of young adults living at home*

Consider the hypothesis that 'one in four UK young adults aged between 20 and 34 years still live with their parents' from the Introduction. The first thing that we need to do is to reinterpret this hypothesis in terms of a parameter. The words 'one in four' suggest that an appropriate parameter would be p, the proportion of young adults between 20 and 34 years who still live with their parents. So this hypothesis can be reinterpreted as testing whether $p = 0.25$.

The null hypothesis representing 'no difference' therefore suggests specifying the null hypothesis to be

$$H_0 : p = 0.25.$$

You can now try specifying the null hypothesis for some tests yourself.

Activity 1 *Null hypothesis: driving theory test pass rates*

In 1996, a driving theory test was introduced in the UK which drivers are required to pass before they can take the driving practical test. The pass rate for the theory test was quite high, peaking with 70.6% passing the test nationally in August 2008. However, following a series of measures introduced between 2007 and 2012 to make the test more difficult, the pass rate fell, and on 11 September 2014, a *BBC News* article had the headline: 'Harder theory test leads to falling pass rate'.

The theory test is taken at a number of different test centres across the UK. The data that are available are the pass rates for a sample of these centres over the period April 2014–March 2015. Over the period April 2013–March 2014, the overall pass rate nationally was 51.6%. (Because this is the national pass rate, it can be considered to be a known, population, value.) Specify a null hypothesis for testing whether the average pass rate nationally over the period April 2014–March 2015 is lower than the national pass rate for the same period the previous year.

An overall pass rate and an average pass rate are not the same thing, but should be good approximations to one another in this context. Comparing the two is a consequence of the form of data that we use.

Activity 2 *Null hypothesis: blueberries and blood pressure*

Consider once again the claim that 'blueberries lower blood pressure in menopausal women' mentioned in the Introduction. A study was conducted to investigate this issue. (Source: Johnson, S.A. et al. (2015) 'Daily blueberry consumption improves blood pressure and arterial stiffness in postmenopausal women with pre- and stage 1-hypertension: a randomized, double-blind, placebo-controlled clinical trial', *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 3, pp. 369–77.) Forty post-menopausal women aged 45–65 were recruited to the study. Every day for 8 weeks, 20 of the women were given 22 g of freeze-dried blueberry powder, while the rest were given a different powder with matching nutrients to the blueberry powder. None of the women knew which powder they were taking.

Blood pressure (BP) readings have two numbers: one for systolic BP and one for diastolic BP. The baseline mean systolic and diastolic BP readings for all post-menopausal women of interest in this study can be taken to be 138 mm Hg and 80 mm Hg, respectively.

The *NHS Choices* website defines systolic pressure as the pressure when your heart pushes blood out and diastolic pressure as the pressure when your heart rests between beats. Both are measured in units of millimetres of mercury (mm Hg).

(a) Write down an appropriate null hypothesis for testing whether the mean systolic BP for the women who took blueberry powder is lower than the mean systolic BP for all post-menopausal women.

(b) Write down an appropriate null hypothesis for testing whether the mean diastolic BP for the women who took blueberry powder is lower than the mean diastolic BP for all post-menopausal women.

The choice of alternative hypothesis depends on the context and purpose of the statistical test. There are three possible choices for H_1:

- $H_1 : \theta \neq \theta_0$
- $H_1 : \theta > \theta_0$
- $H_1 : \theta < \theta_0$.

If it is important for the test to detect only whether θ is greater than θ_0, then an alternative hypothesis of the form $H_1 : \theta > \theta_0$ would be appropriate, whereas if it is important for the test to detect only whether θ is less than θ_0, then an alternative hypothesis of the form $H_1 : \theta < \theta_0$ would be appropriate. On the other hand, if any difference (both greater than and less than θ_0) needs to be detected, then an alternative hypothesis of the form $H_1 : \theta \neq \theta_0$ would be appropriate.

Example 3 *Alternative hypothesis: festive weight gain*

In Example 1, the null hypothesis $H_0 : \mu = 2.3$ kg was specified for testing the hypothesis that μ could be around 2.3 kg. For this test, we would like to detect any difference, either greater than or less than 2.3 kg, so a suitable alternative hypothesis would be

$$H_1 : \mu \neq 2.3 \text{ kg}.$$

Example 4 *Alternative hypothesis: blueberries and systolic blood pressure*

In Activity 2(a), the null hypothesis $H_0 : \mu_S = 138$ mm Hg was specified for testing whether the mean systolic BP was lower for women taking 22 g of freeze-dried blueberry powder every day for 8 weeks. For this test, we would like to detect whether μ_S is less than 138 mm Hg, so a suitable alternative hypothesis would be

$$H_1 : \mu_S < 138 \text{ mm Hg}.$$

Now try specifying some alternative hypotheses for yourself.

Activity 3 *Alternative hypothesis: blueberries and diastolic blood pressure*

In Activity 2(b), the null hypothesis $H_0 : \mu_D = 80$ mm Hg was specified for testing whether the mean diastolic BP was lower for women taking 22 g of freeze-dried blueberry powder every day for 8 weeks. Specify a suitable alternative hypothesis for this test.

Activity 4 *Alternative hypothesis: driving theory test pass rates*

In Activity 1, the null hypothesis $H_0 : \mu = 51.6\%$ was proposed for testing whether the average pass rate nationally for the driving theory test over the period April 2014–March 2015 is lower than the national pass rate for the same period the previous year. Specify a suitable alternative hypothesis for this test.

Activity 5 *Alternative hypothesis: proportion of young adults living at home*

In Example 2, the null hypothesis $H_0 : p = 0.25$ was proposed for testing the hypothesis that one in four young adults aged between 20 and 34 years still live with their parents. Specify a suitable alternative hypothesis for this test.

Exercise on Section 1

Exercise 1 *Driving theory test pass rates for males and females*

In addition to the data on driving theory test pass rates for different centres over the period April 2014–March 2015 considered in Activities 1 and 4, data are also available for the separate pass rates for males and for females for each of the test centres over the same period.

(a) Over the period April 2013–March 2014, the national pass rate for females was 54.7%. Specify null and alternative hypotheses for testing whether the average pass rate nationally for females for the period April 2014–March 2015 is different to the national pass rate for females for the same period the previous year.

(b) Over the period April 2013–March 2014, the national pass rate for males was 48.8%. Specify null and alternative hypotheses for testing whether the average pass rate nationally for males for the period April 2014–March 2015 is less than the national pass rate for males for the same period the previous year.

2 Testing hypotheses: the main ideas

In order to introduce the main ideas of hypothesis testing, we will once again consider the specific problem of testing the claim that the average weight gain over the festive period is 2.3 kg.

The units of measurement, 'kg', will be assumed, but generally not written, in what follows.

In Examples 1 and 3, the null and alternative hypotheses for this problem were specified as

$$H_0 : \mu = 2.3, \quad H_1 : \mu \neq 2.3.$$

Now some data are required to test these hypotheses.

The US 'festive period' was considered to start on Thanksgiving (towards the end of November) and finish on New Year's Day.

In a study in the USA to investigate weight gain over the festive period, 195 adults were weighed in mid-November and again in early to mid-January. (Source: Yanovski, J.A. et al. (2000) 'A prospective study of holiday weight gain', *New England Journal of Medicine*, vol. 342, no. 12, pp. 861–7.) To decrease the chance that the people being weighed might subconsciously attempt to change their weight gain between measurements, no one was told that investigating their weight gain was the primary purpose of the study (it was masked by taking several different measurements). The sample mean weight gain for the 195 adults during the festive period was 0.37 kg, and the sample standard deviation was 1.52 kg.

Christmas is in this festive period

On the face of it, the observed sample mean weight gain of 0.37 seems to be quite a lot less than the hypothesised value of the population mean weight gain, $H_0 : \mu = 2.3$, which may lead you to doubt that H_0 is true. But is it possible that H_0 is true so that μ *is* 2.3 and an observed sample mean value of 0.37 could have arisen by chance? To answer this question, we need to consider what the distribution of the sample mean is if H_0 is true.

If we assume that H_0 is true, then the festive weight gains, $X_1, X_2, \ldots, X_{195}$, should have mean μ equal to 2.3 with some variance σ^2. In this case, since the sample size, $n = 195$, is large, by the Central Limit Theorem the sample mean, $\overline{X}$, should be approximately normally distributed with mean $\mu = 2.3$ and variance $\sigma^2/n = \sigma^2/195$. We do not know the value of σ^2, although we do have an estimate of it from the sample: this estimate is $s^2 = 1.52^2$. So, *if* H_0 *is true*, then an approximate distribution for $\overline{X}$ is

See Subsection 6.2 of Unit 6.

$$\overline{X} \approx N\left(2.3, \frac{1.52^2}{195}\right) \simeq N(2.3, 0.0118).$$

So if H_0 is true, then the observed sample mean, $\overline{x} = 0.37$, will have arisen from this distribution. The next step is to decide whether or not it looks likely that the observed value of $\overline{x}$ could have arisen from $N(2.3, 0.0118)$.

A plot of the p.d.f. for $N(2.3, 0.0118)$ is given in Figure 1.

A value of $\overline{x} = 0.37$ is so far into the lower tail of the p.d.f. for $N(2.3, 0.0118)$ that it is not even marked on the horizontal axis in Figure 1! So, clearly, an observed value of 0.37 really is *not* likely to have been observed from this distribution, which casts doubt on H_0 being true. The data therefore suggest that we should reject the null hypothesis that $\mu = 2.3$ and conclude that the mean festive weight gain is not equal to 2.3 kg.

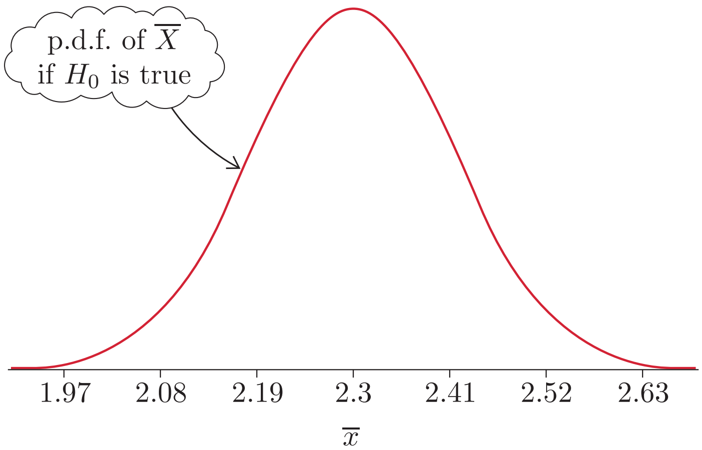


Figure 1 Plot of the p.d.f. for $N(2.3, 0.0118)$

The word 'reject' is one of several synonyms for 'dismiss' or 'discard', but is the one always used by statisticians as a technical term for finding evidence against the null hypothesis in the context of hypothesis testing.

This test of the truth of the null hypothesis was based on using the observed value of a sample statistic, the sample mean, and the distribution of that sample statistic assuming that the null hypothesis is true. The sample statistic on which a test is based is called the **test statistic**, and its distribution assuming that the null hypothesis is true is called the **null distribution**.

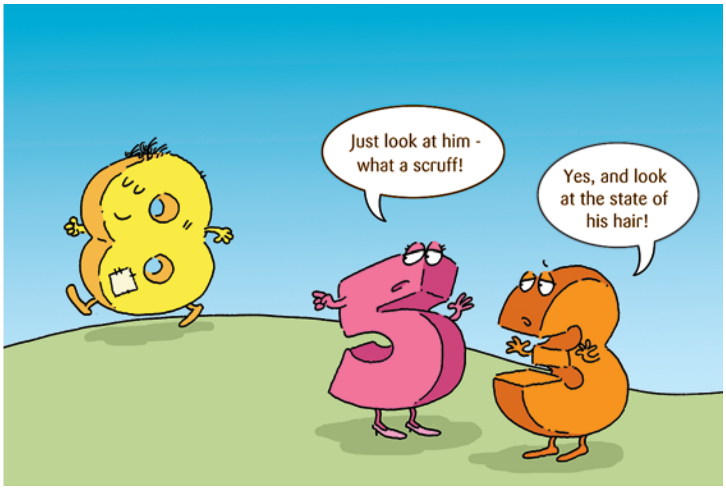
Looking again at Figure 1, if $\overline{x}$ had been 2.0 kg, say, rather than 0.37 kg, then such an observed value would be more in line with H_0 being true, but it is still a rather unlikely value to have occurred under H_0 since the value $\overline{x} = 2.0$ lies quite far into the lower tail of the null distribution, $N(2.3, 0.0118)$. So how might we decide which values of $\overline{x}$ will be considered to be unlikely enough to reject H_0?

Since the alternative hypothesis is $\mu \neq 2.3$, we would like to detect whether μ seems to be smaller than 2.3, as well as whether μ seems to be larger than 2.3. To do this, we find two values, c_1 and c_2, so that if H_0 is true,

$$P(\overline{X} \leq c_1) = P(\overline{X} \geq c_2) = \frac{\alpha}{2},$$

where α is small. That is, we find c_1 and c_2 so that if H_0 is true, the probability α that the sample mean $\overline{X}$ is as or more extreme than either c_1 and c_2 is small. Then, we reject H_0 if the observed value of $\overline{x}$ is rather surprising under the null hypothesis, that is, if it is either less than or equal to c_1, or greater than or equal to c_2. The values c_1 and c_2 are called **critical values**, and the set of values of $\overline{x}$ which would lead us to reject H_0 is known as the **rejection region**. This is illustrated in Figure 2 (overleaf).

The rejection region is also often called the critical region.



Critical values

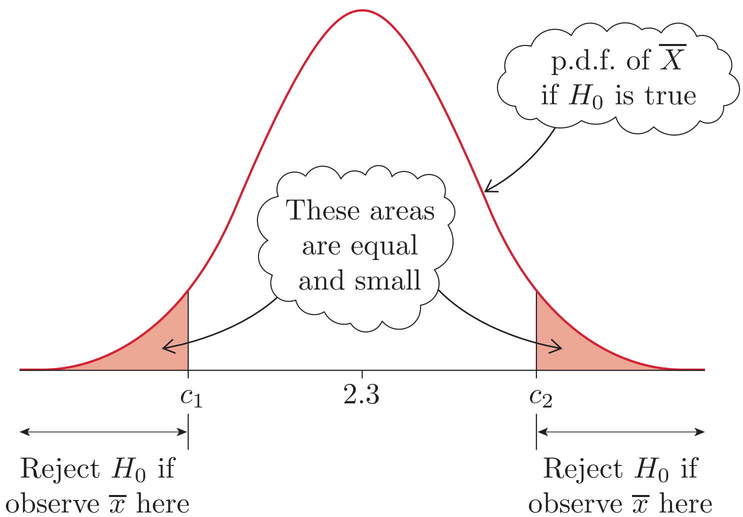


Figure 2 Plot of the p.d.f. for $N(2.3, 0.0118)$ with critical values and rejection region

Conventionally, c_1 and c_2 are chosen so that α is 0.01, 0.05 or 0.1. The value $100\alpha\%$ is then called the **significance level** of the test. For example, when $\alpha = 0.05$, the significance level is 5%.

In the weight gains example, if H_0 is true, then $\overline{X} \approx N(2.3, 0.0118)$. So when $\alpha = 0.05$ (i.e. the significance level is 5%), the critical values c_1 and c_2 are such that

$$P(\overline{X} \leq c_1) = 0.025 \quad \text{and} \quad P(\overline{X} \geq c_2) = 0.025.$$

To find the values of c_1 and c_2, we need to standardise $\overline{X}$ (see Subsection 4.4 of Unit 6). Standardising $\overline{X}$'s distribution,

$$Z = \frac{\overline{X} - 2.3}{\sqrt{0.0118}} \approx N(0,1).$$

This means that

$$P(\overline{X} \leq c_1) = P\left(Z \leq \frac{c_1 - 2.3}{\sqrt{0.0118}}\right) = 0.025,$$

so, following Examples 15 and 16 in Unit 6,

$$\frac{c_1 - 2.3}{\sqrt{0.0118}} = q_{0.025} = -q_{1-0.025} = -q_{0.975}.$$

It follows that, from the table of quantiles of the standard normal distribution, Table 6 in Unit 6, which is reproduced in the Handbook,

$$\frac{c_1 - 2.3}{\sqrt{0.0118}} = -1.960,$$

so

$$c_1 = 2.3 - 1.960\sqrt{0.0118} \simeq 2.087.$$

Similarly,

$$P(\overline{X} \geq c_2) = P\left(Z \geq \frac{c_2 - 2.3}{\sqrt{0.0118}}\right) = 0.025,$$

so, following Example 11 in Unit 6,

$$P\left(Z \geq \frac{c_2 - 2.3}{\sqrt{0.0118}}\right) = 1 - P\left(Z \leq \frac{c_2 - 2.3}{\sqrt{0.0118}}\right) = 0.025,$$

so

$$P\left(Z \leq \frac{c_2 - 2.3}{\sqrt{0.0118}}\right) = 0.975$$

and

$$\frac{c_2 - 2.3}{\sqrt{0.0118}} = q_{0.975}.$$

Then, from the table of standard normal quantiles,

$$\frac{c_2 - 2.3}{\sqrt{0.0118}} = 1.960,$$

so

$$c_2 = 2.3 + 1.960\sqrt{0.0118} \simeq 2.513.$$

The rejection region is therefore defined to be all values of $\overline{x}$ which are less than or equal to 2.087, or greater than or equal to 2.513.

Since 2.0 is less than 2.087 (just!), and so is in the rejection region, an observed value of $\overline{x} = 2.0$ would lead us to reject H_0 at the 5% significance level. This is illustrated in Figure 3 (overleaf).

For the real data, $\overline{x} = 0.37$, which is well into the rejection region, confirming (more formally) our earlier conclusion that the data suggest that the mean festive weight gain is not equal to 2.3 kg. What's more, the fact that the observed sample mean is so much lower than 2.3 kg would suggest that the mean festive weight gain is in fact lower than 2.3 kg.

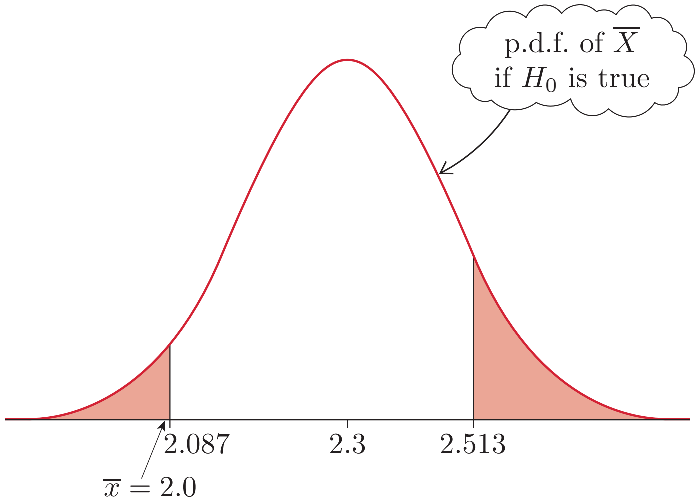


Figure 3 Null distribution with the critical values marked, together with the observed test statistic $\overline{x}$

This specific example considering mean festive weight gain has introduced the main steps in carrying out a hypothesis test, which are summarised in the following box.

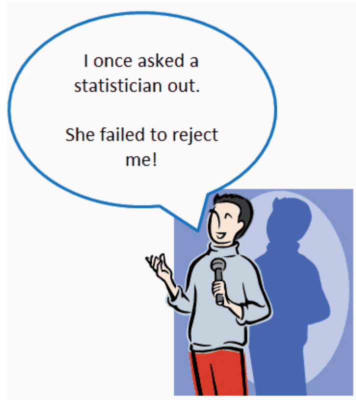
The main steps in a hypothesis test

- Set up the **null hypothesis**, H_0, and the **alternative hypothesis**, H_1.
- Obtain some sample data and summarise these in the **test statistic**.
- Obtain the **null distribution**. This is the distribution of the test statistic under the assumption that H_0 is true.
- Decide on the **significance level** for the test. This is usually one of 1%, 5% or 10%.
- Calculate the **critical values** for the significance level, and hence the **rejection region** for the test.
- Make one of two possible decisions:
  - Reject H_0 if the test statistic lies in the rejection region.
  - Do not reject H_0 if the test statistic does not lie in the rejection region.
- State the conclusion of the test in non-technical language.

Such a hypothesis test is often called a 'fixed level' hypothesis test because you have to choose a value for the significance level.

Notice that the decision 'do not reject H_0' is *not* the same as deciding that H_0 is true. Just because the observed data are quite likely to have arisen if H_0 is true, this does not mean that H_0 *is* true – it means only that H_0 *might* be true and, in particular, that we do not have sufficient evidence on the basis of these data to reject it. For example, if the observed data were very likely if $H_0 : \mu = 2.3$ kg is true, then it might also be the case that the observed data are also very likely if μ was in fact 2.2 kg or 2.4 kg, so that,

given the data, these would also be possible values for μ. As such, it would be incorrect for us to conclude that $H_0 : \mu = 2.3$ kg is true. (It is also inappropriate to say that we 'accept' H_0 – as you will sometimes see written – if H_0 is not rejected; again, all we can say is that we do not reject H_0.)



You can hopefully see the main idea of hypothesis testing quite clearly now. We essentially ask: Could the data we observe have arisen due to random variation if the null hypothesis is true? If the answer is 'no', we reject H_0 in favour of the alternative H_1; if the answer is 'yes', we do not reject H_0 – it remains plausible, given the data.

All that said, you have been given a lot of information over the last few pages and it would be useful for you to pause now and digest the information. So in the next activity you are asked to summarise each of the main steps of the festive weight gain hypothesis test that was carried out above.

Activity 6 *Summarising the test of festive weight gain*

Considering the test of festive weight gain which has just been carried out:

(a) State the null and alternative hypotheses for the test.

(b) Summarise the sample data from the US study that were used for the test, and identify the test statistic that was used.

(c) State the null distribution of the test statistic.

(d) State what significance level was used for the test.

(e) State the critical values and rejection region for the test.

(f) State whether H_0 should be rejected or not, given the observed value of the test statistic and the rejection region.

(g) State the conclusion of the test in non-technical language.

In the example just described, $\overline{X}$ was used as the test statistic with observed value $\overline{x} = 0.37$ kg, and the null distribution was the distribution for $\overline{X}$ assuming that H_0 is true: $\overline{X} \approx N(2.3, 0.0118)$. The null distribution was then standardised when calculating each of the critical values c_1 and c_2. An alternative, but in fact equivalent, test could have been specified by using the standardised value of $\overline{X}$ when the null hypothesis is true,

$$Z = \frac{\overline{X} - 2.3}{\sqrt{0.0118}},$$

as the test statistic. In this case, the observed value of the test statistic is

$$z = \frac{\overline{x} - 2.3}{\sqrt{0.0118}} = \frac{0.37 - 2.3}{\sqrt{0.0118}} \simeq -17.77,$$

and the null distribution is (approximately) $N(0, 1)$.

So why might we want to use this alternative test statistic? Well, because the null distribution is $N(0,1)$, it is very easy to calculate critical values for the test because c_1 and c_2 are such that

$$P(Z \leq c_1) = \frac{\alpha}{2} \quad \text{and} \quad P(Z \geq c_2) = \frac{\alpha}{2},$$

so

$$c_1 = q_{\alpha/2} \quad \text{and} \quad c_2 = q_{1-(\alpha/2)}$$

(see Figure 4).

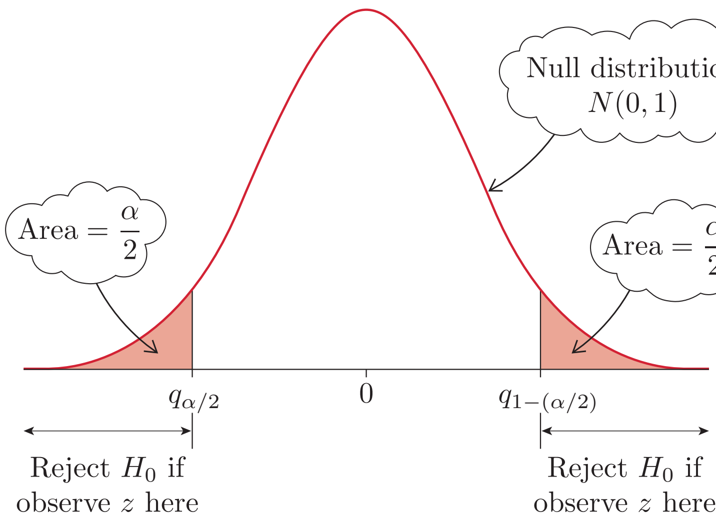


Figure 4 Plot of the null distribution $N(0,1)$, with critical values and rejection region marked

So, for example, using a 5% significance level, from the table of standard normal quantiles and Example 16 in Unit 6,

$$c_1 = q_{0.025} = -q_{0.975} = -1.960 \quad \text{and} \quad c_2 = q_{0.975} = 1.960.$$

The rejection region is therefore all values of z which are less than or equal to -1.960 or greater than or equal to 1.960. The observed value of z is -17.77, which is a long way below -1.960, leading to the rejection of H_0 and confirmation of the result of the test when using the test statistic $\overline{X}$ instead of Z.

You will use Z, the standardised value of $\overline{X}$, as the test statistic to test different hypotheses regarding the mean festive weight gain in the following activity.

Food for a different festive season, Eid ul-Fitr, at the conclusion of Ramadan

Activity 7 *Different hypotheses for festive weight gain*

Suppose that instead of testing whether the mean festive weight gain is 2.3 kg, you wish to test whether the mean festive weight gain is 0.55 kg. In this activity, you will test this hypothesis using the observed data $\overline{x} = 0.37$ kg and $s = 1.52$ kg from the US study.

(a) State the (new) null and alternative hypotheses for the test.

(b) State the approximate normal distribution for $\overline{X}$ assuming that H_0 is true.

(c) Hence specify the observed value of a test statistic which, if H_0 is true, can be assumed to be an observation from $N(0,1)$.

(d) Using a 5% significance level, what are the critical values and the rejection region for the test?

(e) State whether or not H_0 should be rejected.

(f) State the conclusion of the test in non-technical language.

So far in this section we have considered tests only for hypotheses for which the alternative hypothesis is of the form $H_1: \theta \neq \theta_0$. However, in Section 1 there were alternative hypotheses of the form $H_1: \theta < \theta_0$ and $H_1: \theta > \theta_0$. How would these hypotheses be tested? The only real difference for testing these latter alternative hypotheses is in the rejection region for the test and in the conclusion of the test.

When testing the festive weight gain hypotheses

$$H_0: \mu = 2.3, \quad H_1: \mu \neq 2.3,$$

H_0 is rejected if the test statistic lies in either tail of the null distribution because we would like to be able to detect any difference from the value 2.3, both low and high (see Figure 2). Now suppose that we're interested in testing only whether the mean festive weight gain is *lower* than 2.3 kg, so that the hypotheses are

$$H_0: \mu = 2.3, \quad H_1: \mu < 2.3.$$

This time we are interested only in detecting low values of the mean which might support the alternative hypothesis. Thus we'll reject H_0 only if we observe the test statistic in the *lower* tail of the null distribution. So, using Z as our test statistic (and hence the null distribution is $N(0,1)$), with a 5% significance level we find a *single* critical value, c_1, such that

$$P(Z \leq c_1) = 0.05,$$

so that, from the table of standard normal quantiles and Example 16 in Unit 6,

$$c_1 = q_{0.05} = -q_{0.95} = -1.645.$$

The rejection region is therefore all values of z less than or equal to -1.645. This is illustrated in Figure 5 (overleaf).

The observed value of z is still -17.77, which is indeed less than -1.645. We therefore reject H_0 at the 5% significance level and conclude that the data suggest that the mean festive weight gain is less than 2.3 kg. Notice that the conclusion reflects the form of the alternative hypothesis.

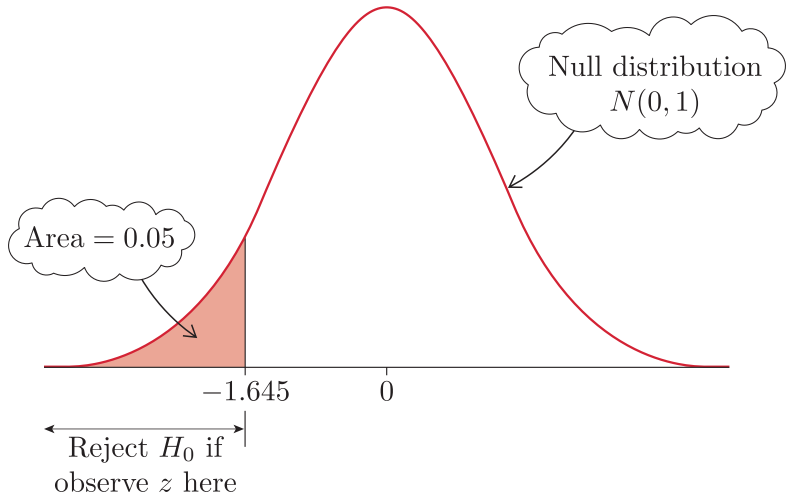


Figure 5 Plot of the null distribution with rejection region for testing $H_0 : \mu = 2.3$, $H_1 : \mu < 2.3$

Activity 8 *Festive weight gain lower than 0.55 kg?*

In Activity 7, you tested whether the mean festive weight gain is 0.55 kg, with the alternative hypothesis being that the mean festive weight gain is different from 0.55 kg. In this activity, you will test whether the mean festive weight gain is 0.55 kg when the alternative hypothesis is that the mean festive weight gain is *lower* than 0.55 kg. As in Activity 7, you will use the observed data $\overline{x} = 0.37$ kg and $s = 1.52$ kg from the US study.

(a) State the (new) null and alternative hypotheses for the test.

(b) Calculate the observed value of the test statistic Z which, if H_0 is true, is approximately distributed as $N(0, 1)$.

(c) Using a 5% significance level, what are the critical value and the rejection region for the test?

(d) State whether or not H_0 should be rejected.

(e) State the conclusion of the test in non-technical language.

A two-tailed lizard!

Because the rejection region of any test with alternative hypothesis of the form $\theta \neq \theta_0$ contains values in both tails of the null distribution, any test with an alternative hypothesis of this form is called a **two-tailed test** or a **two-sided test**.

On the other hand, the rejection region of a test with $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$ contains values in only one tail of the null distribution, so is called a **one-tailed test** or a **one-sided test**.

3 Testing hypotheses: some common tests

In this section, we consider some commonly used tests for the mean of a population and for an unknown proportion.

3.1 Testing the mean of a population

Consider testing the hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

where μ is the mean of some population, and μ_0 is some specific value.

Recall that μ_0 is the value of μ which represents 'no difference'.

The z-test

In Section 2, hypotheses of this form were tested regarding μ, the mean festive weight gain. There, two test statistics were considered: $\overline{X}$ and Z, the standardised value of $\overline{X}$ when H_0 is true (using the sample variance as an estimate of the population variance). Because the sample size, n, was large, the Central Limit Theorem meant that the null distributions for both of these test statistics were approximately normal. In the case of Z the null distribution is $N(0, 1)$, which makes it particularly easy to find the critical values, and hence the rejection region, of the test. Using Z as a test statistic for testing hypotheses regarding a population mean μ when the sample size is large, results in a commonly used test called the ***z*-test**.

(This is often called the 'one-sample z-test', 'one-sample' referring to the fact that we are considering a single mean and sample here; a 'two-sample z-test', concerned with differences between parameter values on the basis of two independent samples of data, is also popular, but there is no room to consider it in this module.)

More formally, the z-test uses the fact that if $X_1, X_2, \ldots, X_n$ are n independent random variables from a population with mean μ and variance σ^2, then for large n,

Rule of thumb from Unit 6: n is 'large' if it is at least 25.

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

Now, if H_0 is true, then μ is the value μ_0, which means that

$$\overline{X} \approx N\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Further, since n is large, the sample variance, s^2, will be close to the population variance, σ^2, so

Rule of thumb from Unit 8: even when σ^2 is estimated by s^2, n is still 'large' if it is at least 25.

$$\overline{X} \approx N\left(\mu_0, \frac{s^2}{n}\right).$$

Thus if H_0 is true, the standardised value of $\overline{X}$ is

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \approx N(0, 1).$$

All wired up for a zzzzz-sleep test ...

The z-test therefore uses this Z as the test statistic, and $N(0,1)$ is the null distribution. The critical values for the test are then the appropriate standard normal quantiles obtained from Table 6 of Unit 6, and reproduced in the Handbook. For example, for a 5% significance level,

$$c_1 = q_{0.025} = -q_{0.975} = -1.960 \quad \text{and} \quad c_2 = q_{0.975} = 1.960,$$

so the rejection region is values of z such that $z \leq -1.960$ or $z \geq 1.960$.

Recall that the only real differences between two-sided and one-sided tests are the critical values, the consequent rejection region, and the conclusion of the test. When the alternative hypothesis has the form $H_1 : \mu < \mu_0$, the corresponding critical value for a 5% significance level is

$$c_1 = q_{0.05} = -q_{0.95} = -1.645,$$

and when the alternative hypothesis has the form $H_1 : \mu > \mu_0$, the corresponding critical value for a 5% significance level is

$$c_2 = q_{0.95} = 1.645.$$

Example 5 *Driving theory test pass rates*

Activities 1 and 4 considered the problem of testing whether the average pass rate for the driving theory test nationally across all UK centres over the period April 2014–March 2015 was lower than the overall national pass rate for the same period the previous year. The following hypotheses were specified:

$$H_0 : \mu = 51.6\%, \quad H_1 : \mu < 51.6\%.$$

The sample mean and sample standard deviation of the pass rates over the period April 2014–March 2015 for $n = 137$ centres were 51.261% and 2.344%, respectively. Since n is large, a z-test can be used with observed test statistic

Data extracted from https://www.gov.uk/government/statistical-data-sets/car-theory-test-data-by-test-centre; more complete data are not used here in order to mimic common situations in which hypothesis tests are required.

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{51.261 - 51.6}{2.344/\sqrt{137}} \simeq -1.693.$$

Since the test is one-sided, with a 5% significance level the critical value is $c_1 = -1.645$ and the rejection region is all values of z less than or equal to -1.645. Then since $-1.693 < -1.645$, the observed value of the test statistic lies in the rejection region, so H_0 should be rejected at the 5% significance level. We conclude that the data suggest that the mean pass rate nationally over the period April 2014–March 2015 is lower than the national pass rate the previous year.

Activity 9 *Driving theory test pass rates for females and males*

(a) In Exercise 1(a), null and alternative hypotheses were specified for testing whether μ_F, the average driving theory test pass rate for females nationally over the period April 2014–March 2015, is different

to the national pass rate for females for the same period the previous year. The hypotheses were

$$H_0 : \mu_F = 54.7\%, \quad H_1 : \mu_F \neq 54.7\%.$$

The sample mean and sample standard deviation of the pass rates for females over the period April 2014–March 2015 for $n = 137$ centres were 54.166% and 2.864%, respectively.

(i) Since n is large, a z-test will be used to test the hypotheses. Calculate the observed value of the test statistic for this test.

(ii) Obtain the critical values and rejection region associated with the test at the 5% significance level.

(iii) What do you conclude about whether the average driving theory test pass rate for females nationally over the period April 2014–March 2015 is different to the national pass rate for females for the same period the previous year?

(b) In Exercise 1(b), null and alternative hypotheses were specified for testing whether μ_M, the average driving theory test pass rate for males nationally over the period April 2014–March 2015, is lower than the national pass rate for males for the same period the previous year. The hypotheses were

$$H_0 : \mu_M = 48.8\%, \quad H_1 : \mu_M < 48.8\%.$$

The sample mean and sample standard deviation of the pass rates for males over the period April 2014–March 2015 for $n = 137$ centres were 48.683% and 2.336%, respectively.

(i) Since n is large, a z-test will be used to test the hypotheses. Calculate the observed value of the test statistic for this test.

(ii) Obtain the critical value and rejection region associated with the test at the 10% significance level.

(iii) What do you conclude about whether the average driving theory test pass rate for males nationally over the period April 2014–March 2015 is lower than the national pass rate for males for the same period the previous year?

(c) Comment on the result of the test of the national pass rate given in Example 5 in the light of the results of the tests in parts (a) and (b).

The t-test

The z-test can be used for testing hypotheses regarding the population mean, regardless of the population distribution. It does, however, require the sample size to be large because it uses the Central Limit Theorem. So can hypotheses about the population mean be tested when the sample size is not large (which by the rule of thumb means that $n < 25$)? The answer to this is ‘yes’ when the population distribution is normal.

Suppose that we wish to test hypotheses regarding the mean of a normal population. In Subsection 4.2 of Unit 8, you saw that for a random sample of size n from a normal population, the random variable is

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

A tea test ...

where $t(n-1)$ denotes the t-distribution with $n-1$ degrees of freedom. So when testing the hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

if the population distribution is normal, then when H_0 is true,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Then T can be used as a test statistic with null distribution $t(n-1)$. Like the z-test, this test is in common use and is called the ***t*-test**.

Again, a two-sample version of this one-sample t-test exists but will not be covered in this module.

Example 6 *Blueberries and systolic blood pressure*

In Activity 2 and Example 4, null and alternative hypotheses were specified for testing whether taking 22 g of freeze-dried blueberry powder every day for 8 weeks lowers μ_S, the mean systolic BP for menopausal women. The hypotheses were

$$H_0 : \mu_S = 138 \text{ mm Hg}, \quad H_1 : \mu_S < 138 \text{ mm Hg}.$$

The sample mean and sample standard deviation of systolic BP for $n = 20$ women taking 22 g of freeze-dried blueberry powder every day for 8 weeks were 131 mm Hg and 17 mm Hg, respectively. Graphical analysis by the study researchers allowed the assumption that the systolic BP measurements follow a normal distribution, so a t-test can be used to test the hypotheses.

For these data, the observed value of the test statistic, T, is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{131 - 138}{17/\sqrt{20}} \simeq -1.841.$$

This is a one-sided test, so using a 5% significance level, the critical value c_1 is the 0.05-quantile of the $t(20-1) = t(19)$ distribution. But because the t-distribution is symmetric about 0, this means that c_1 is the negative of the 0.95-quantile. From Table 5 in Unit 8, which can also be found in the Handbook, the 0.95-quantile of $t(19)$ is 1.729, so $c_1 = -1.729$ and the rejection region is all values of t such that $t \leq -1.729$.

Since $-1.841 < -1.729$, the observed value of t is in the rejection region, so we reject H_0 at the 5% significance level. We conclude that the data suggest that taking 22 g of blueberry powder each day for 8 weeks lowers systolic BP in menopausal women.

Activity 10 *Blueberries and diastolic blood pressure*

In Activities 2 and 3, null and alternative hypotheses were specified for testing whether taking 22 g of freeze-dried blueberry powder every day for 8 weeks lowers μ_D, the mean diastolic BP for menopausal women. The hypotheses were

$$H_0: \mu_D = 80 \text{ mm Hg}, \quad H_1: \mu_D < 80 \text{ mm Hg}.$$

The sample mean and sample standard deviation of diastolic BP for $n = 20$ women taking 22 g of freeze-dried blueberry powder every day for 8 weeks were 75 mm Hg and 9 mm Hg, respectively.

(a) The diastolic BP measurements can be assumed to follow a normal distribution, and a t-test will be used to test the hypotheses. Calculate the observed value of the test statistic for this test.

(b) Obtain the critical value and rejection region associated with the test at the 5% significance level.

(c) What do you conclude about whether taking 22 g of freeze-dried blueberry powder every day for 8 weeks lowers the mean diastolic BP for menopausal women?

For testing population means, data are sometimes collected as paired observations. For example, to assess the effect of a drug designed to lower cholesterol, the cholesterol for a set of patients could be measured both before and after taking the drug so that there are two measurements for each patient: the 'before' measurement and the 'after' measurement. In this case, the two observations for each patient are said to be *paired*, and the two observations will be related because they are measured on the same individual (in contrast to measurements on two different individuals being independent).

Such data are also called 'matched pairs'.

For such data, the differences between the paired observations are often the focus of interest. These differences can be treated as a single sample (which happens to consist of differences between two values), and tests concerning the mean differences can be carried out using a (one-sample) z- or t-test as required. For example, for the cholesterol example, we would calculate the differences,

$$d_i = \text{'before' measurement for patient } i - \text{'after' measurement for patient } i,$$

and then treat $d_1, d_2, \ldots, d_n$ as our data. A test of the hypotheses

$$H_0: \mu_D = 0, \quad H_1: \mu_D < 0,$$

where μ_D is the mean difference, would then test whether the drug lowered cholesterol. Testing paired observations using the differences $d_1, d_2, \ldots, d_n$ is essentially no different to testing a single sample of data: you will carry out such a test using Minitab when you work through Chapter 7 of Computer Book B (in Subsection 4.1).

Peared observations? Measurements on a pair of pears growing together on a tree would not be independent.

Tests for a population mean are summarised in the following box.

Testing a population mean

There are two commonly used tests for testing the null hypothesis

$$H_0 : \mu = \mu_0$$

against one of the alternative hypotheses

$$H_1 : \mu \neq \mu_0 \quad \text{or} \quad H_1 : \mu < \mu_0 \quad \text{or} \quad H_1 : \mu > \mu_0.$$

***z*-test**

- Can be used *whatever the underlying distribution* when *the sample size is large* ($n \geq 25$).
- The test statistic is
$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$
- The null distribution is $N(0, 1)$.
- Critical values are found from the $N(0, 1)$ quantile table.

***t*-test**

- Can be used for a *normal population* for *any sample size*.
- The test statistic is
$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$
- The null distribution is $t(n - 1)$.
- Critical values are found from the $t(n - 1)$ quantile table.

3.2 Testing a proportion with a large sample

Suppose that random variable X follows a binomial distribution, $B(n, p)$, and we wish to test the null hypothesis

$$H_0 : p = p_0,$$

for some specific proportion p_0, against one of the alternative hypotheses

$$H_1 : p \neq p_0 \quad \text{or} \quad H_1 : p < p_0 \quad \text{or} \quad H_1 : p > p_0.$$

Recall from Subsection 3.2 of Unit 8 that when n is large and both np and $n(1 - p)$ are greater than or equal to 5,

$$\widehat{p} = \frac{X}{n} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

So *when H_0 is true*,

$$\widehat{p} = \frac{X}{n} \approx N\left(p_0, \frac{p_0(1-p_0)}{n}\right).$$

Let Z_p be the standardised value of $\widehat{p}$. Then

$$Z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1),$$

and Z_p can be used as a test statistic for testing the hypotheses, with null distribution $N(0,1)$. Notice how this is similar to the z-test in that a large sample size is required, the test statistic is a standardised variable calculated assuming that H_0 is true, and the null distribution is $N(0,1)$.

Example 7 *Young adults living with parents in Wales*

The Labour Force Survey published in January 2014 by the UK's Office for National Statistics estimated that 25% of all UK young adults aged between 20 and 34 years lived with their parents in 2013. This is a very large survey, so we will assume the proportion 0.25 to be the true proportion of all UK young adults aged between 20 and 34 years who were living with their parents in 2013.

Let p_W denote the proportion of young adults aged 20–34 in Wales who lived with their parents in 2013. The survey found that 68 out of 254 young adults questioned were living with their parents in Wales in 2013. We will use these data to test the hypotheses

$$H_0 : p_W = 0.25, \quad H_1 : p_W \neq 0.25.$$

The sample size $n = 254$ is large, so Z_p is an appropriate test statistic with null distribution $N(0,1)$. So

$$z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{68}{254} - 0.25}{\sqrt{\frac{0.25\times 0.75}{254}}} \simeq 0.652.$$

The test is two-sided, so using a 10% significance level, the critical values (from the table of normal quantiles) are

$$c_1 = -1.645, \quad c_2 = 1.645,$$

and the rejection region is values of z_p such that $z_p \leq -1.645$ or $z_p \geq 1.645$.

Since $-1.645 < 0.652 < 1.645$, the observed value of z_p is not in the rejection region. Thus there is no evidence to reject H_0 at the 10% significance level, and we conclude that there is no evidence to suggest that the proportion of young adults aged 20–34 in Wales is different from 0.25.

Activity 11 *Young adults living with parents in Northern Ireland*

Following Example 7, the Labour Force Survey also found that of the 307 young adults aged 20–34 surveyed in Northern Ireland, 111 lived with their parents in 2013. Letting p_{NI} denote the proportion of young adults aged 20–34 years in Northern Ireland who lived with their parents in 2013, test the hypotheses

$$H_0 : p_{NI} = 0.25, \quad H_1 : p_{NI} \neq 0.25,$$

using a 5% significance level.

The test for a population proportion with a large sample is summarised in the following box.

Testing a population proportion

For proportion p, *when the sample size is large* $(n \geq 25)$, test the null hypothesis

$$H_0 : p = p_0$$

against one of the alternative hypotheses

$$H_1 : p \neq p_0 \quad \text{or} \quad H_1 : p < p_0 \quad \text{or} \quad H_1 : p > p_0,$$

using the test statistic

$$Z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where $\widehat{p} = X/n$. The null distribution is then $N(0,1)$, and critical values are found from the $N(0,1)$ quantile table.

3.3 The link between confidence intervals and hypothesis tests

In Unit 8 you met z-intervals and t-intervals, and in this section we have discussed z-tests and t-tests. It will no doubt come as no surprise to you that there is a direct link between z-intervals and z-tests, and between t-intervals and t-tests. The link between confidence intervals and hypothesis tests will be explored in a particular context in the following screencast.

Screencast 9.1 The link between confidence intervals and hypothesis tests

Exercises on Section 3

Exercise 2 *Insect traps*

A total of 33 insect traps were set out across sand dunes and the numbers of different insects caught in a fixed time were counted. Table 1 gives the number of traps containing various numbers of insects of the taxon *Staphylinoidea*.

The sample mean of the counts is 1.636, and the sample standard deviation is 1.655.

Table 1 *Staphylinoidea* in 33 traps

Count	0	1	2	3	4	5	6	≥ 7
Frequency	10	9	5	5	1	2	1	0

(Source: Gilchrist, W. (1984) *Statistical Modelling*, Chichester, John Wiley and Sons, p. 132)

The devil's coach horse beetle is a member of the *Staphylinoidea* family

Use a z-test and a 5% significance level to test whether μ, the underlying mean number of insects of the taxon *Staphylinoidea* in the traps, is greater than 1.

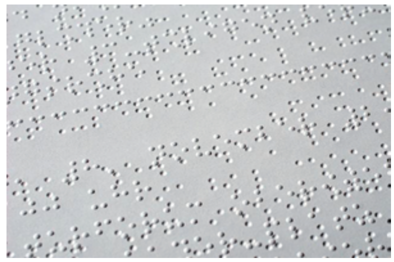
Exercise 3 *Spelling skills of braille readers*

A study in Illinois, USA, tested the spelling skills of a sample of 23 visually impaired children (aged from 6 to 18) who read using braille. Specialist teachers administered a standard 'Test of Written Spelling' (TWS–4), 'norm-referenced' to take account of the children's ages. The tests are designed so that the mean score (at all ages) in the sighted population is 100. Do visually impaired children using braille differ in their average spelling ability relative to sighted children? The data are given in Table 2.

Table 2 Spelling test scores

84	100	83	84	114	96	109	98	98	91	121	111
81	98	110	105	109	88	102	106	114	84	91	

(Source: Clark, C. and Stoner, J.B. (2008) 'An investigation of the spelling skills of braille readers', *Journal of Visual Impairment and Blindness*, vol. 102, no. 9, pp. 553–63)



An example of English braille

The observed sample mean is $\overline{x} = 99.0$ and the sample standard deviation is $s = 11.7$.

Assuming a normal model for the variation in observed test scores, test whether the population mean score of visually impaired children using braille is 100, the same as the mean score in the sighted population, using a 5% level of significance. (Note that departures from the null hypothesis in either direction are of interest; visually impaired children might spell less well than sighted children, or they might be better at spelling.)

Exercise 4 *The proportion of young adults living with parents in Scotland*

Following on from Example 7 and Activity 11, the Labour Force Survey also found that of the 344 young adults aged 20–34 surveyed in Scotland, 86 lived with their parents in 2013.

Letting p_S denote the proportion of young adults aged 20–34 years in Scotland who lived with their parents in 2013, test the hypotheses

$$H_0 : p_S = 0.25, \quad H_1 : p_S \neq 0.25,$$

using a 5% significance level.

4 Significance probabilities: p-values

One question which hasn't been addressed so far is: 'How is the significance level chosen?' It all seems rather arbitrary thus far! This is an important question because it can affect the outcome of the test, as illustrated in the following example.

Example 8 *Driving theory tests: which significance level?*

Example 5 considered testing the hypotheses

$$H_0 : \mu = 51.6\%, \quad H_1 : \mu < 51.6\%,$$

where μ is the mean driving theory test pass rate nationally over the period April 2014–March 2015. The observed value of the test statistic was $z \simeq -1.693$. Using a 5% significance level, the critical value for this one-sided test was $c_1 = -1.645$, which led us to reject H_0 since $-1.693 < -1.645$.

However, if a 1% significance level had been used instead, then the critical value would have been

$$c_1 = q_{0.01} = -q_{0.99} = -2.326,$$

which would mean that H_0 would *not* have been rejected, since $-2.326 < -1.693$.

Both the 1% and 5% critical values, together with the observed value of the test statistic, are marked on a plot of the null distribution in Figure 6.

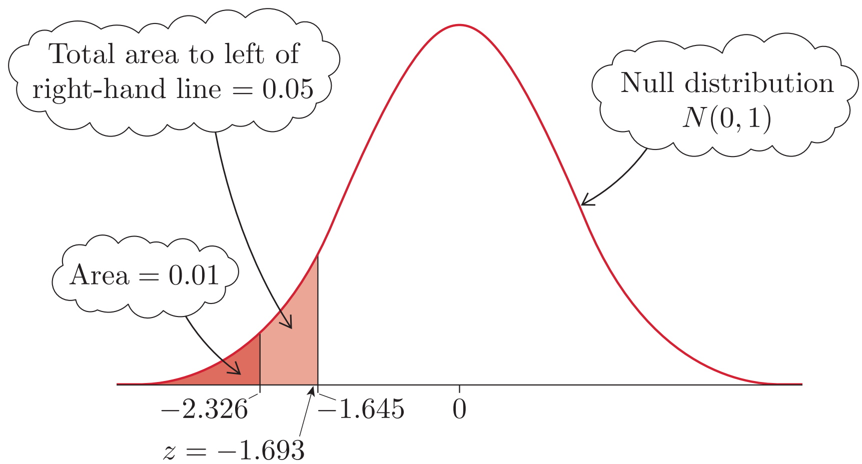


Figure 6 Null distribution $N(0, 1)$ with the 1% critical value -2.326, the 5% critical value -1.645, and the observed value $z = -1.693$ marked

In this section, a different approach to testing hypotheses is considered in which a significance level does not need to be specified for a test. Instead of leading to a stated decision such as 'reject H_0', this alternative approach

to testing hypotheses leads to a number called the **significance probability** which, loosely speaking, describes the extent to which the data provide evidence against the null hypothesis. The general idea is that the *lower* the significance probability, the *more evidence* the data provide against the null hypothesis. The significance probability is denoted by p; it is because of this that it is often referred to as the ***p*-value**.

Do not confuse this use of the letter p with its use to denote the probability of success in a Bernoulli trial. The same letter happens to be standard in both situations. However, note that some books use an upper-case P for p-values.

The use of significance probabilities has become the most common approach for testing hypotheses and, unless you are specifically asked to use critical values and a specific significance level, this is the approach that you should use to perform hypothesis tests.

The idea of a significance probability, or p-value, is to describe the extent to which the data provide evidence against the null hypothesis, rather than to decide whether or not there is sufficient evidence to reject the null hypothesis. This is done by calculating the probability of obtaining a value of the test statistic that is 'at least as extreme as' the observed value of the test statistic when the null hypothesis is true. This is intended as a measure of how likely we are to have observed the data we have observed, or something even more extreme, if the null hypothesis is true. Such a probability, the p-value, will be low when the data are surprising under H_0, so a low p-value corresponds to evidence against the null hypothesis. This is illustrated in Example 9.

Pea value?

Example 9 *Driving theory tests: p-value*

Continuing on from Example 8, the observed value of the test statistic is $z = -1.693$. This is the test statistic for the one-sided test with hypotheses

$$H_0: \mu = 51.6\%, \quad H_1: \mu < 51.6\%,$$

which means that values in the lower tail only will provide evidence against H_0. So any value of z less than -1.693 will be considered to be more extreme than the observed value of the test statistic. Then, since the p-value is the probability of obtaining a value of the test statistic that is 'at least as extreme as' the observed value of the test statistic when the null hypothesis is true, this means that

$$p = P(Z \leq -1.693),$$

where $Z \sim N(0, 1)$ when the null hypothesis is true. So

$$\begin{aligned} p &= P(Z \leq -1.693) = P(Z \geq 1.693) \\ &= 1 - P(Z < 1.693) \simeq 1 - P(Z < 1.69) \\ &= 1 - \Phi(1.69) = 1 - 0.9545 = 0.0455. \end{aligned}$$

(The value -1.693 was rounded to -1.69 to facilitate use of the table of standard normal probabilities used in M248.) The p-value is illustrated in Figure 7 (overleaf).

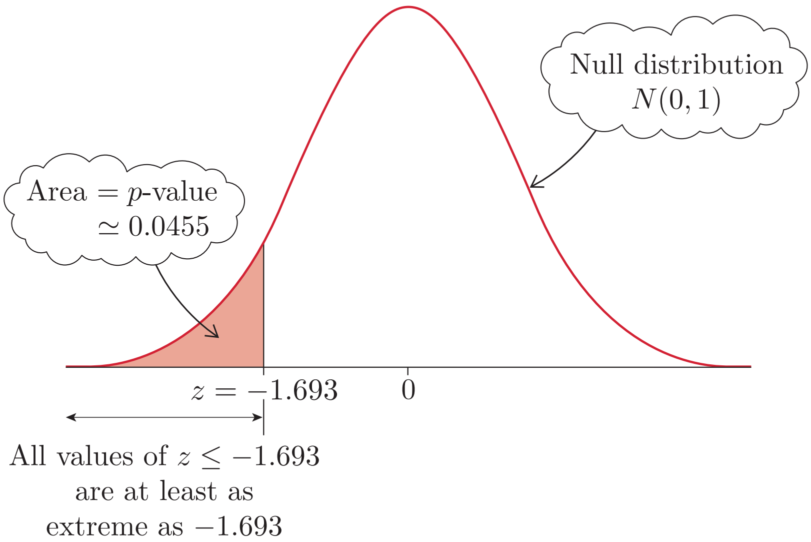


Figure 7 Null distribution $N(0,1)$ with the p-value for the observed value $z = -1.693$ shown

So now that we have calculated a p-value, what do we do with it? What does it tell us about the evidence against the null hypothesis? Well, there aren't any hard and fast rules as to how to interpret p-values, but a rough guide is given in Table 3. Please do be aware that these are *only guidelines* and there will always be situations in which any rule of thumb will be found wanting!

Table 3 Interpreting p-values

p-value	Rough interpretation
$p > 0.10$	little or no evidence against H_0
$0.05 < p \leq 0.10$	weak evidence against H_0
$0.01 < p \leq 0.05$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Example 10 *Driving theory tests: interpretation of p-value*

In Example 9, the p-value was calculated to be 0.0455. Thus $0.01 < p < 0.05$, so by the guidelines in Table 3, the p-value provides moderate evidence against H_0. (Note that in Example 5 we rejected H_0 at the 5% significance level.)

In the next activity you will calculate and interpret a p-value for yourself.

Activity 12 *Driving theory test pass rates for males: p-value*

Activity 9(b) considered testing the hypotheses

$$H_0 : \mu_M = 48.8\%, \quad H_1 : \mu_M < 48.8\%,$$

where μ_M is the average driving theory test pass rate for males nationally over the period April 2014–March 2015.

Given that the observed value of the test statistic is $z = -0.586$ and the null distribution is $N(0,1)$, calculate and interpret the p-value for testing these hypotheses.

So far, we have considered p-values only for one-sided tests. What about the case of a two-sided test? Finding the p-value in a two-sided test is illustrated in the next example.

Example 11 *Driving theory test pass rates for females: p-value*

Activity 9(a) considered the two-sided test with hypotheses

$$H_0 : \mu_F = 54.7\%, \quad H_1 : \mu_F \neq 54.7\%,$$

where μ_F is the average driving theory test pass rate for females nationally over the period April 2014–March 2015. In this case, we're interested in detecting evidence against H_0 as indicated by either a large value of the test statistic, or a small value of the test statistic, that is, values of the test statistic in either tail of the null distribution.

For this test, the observed value of the test statistic was $z = -2.182$, so values of z such that $z \leq -2.182$ would be 'at least as extreme as' the value $z = -2.182$. But this time, observing the test statistic in either the upper tail or the lower tail would cast doubt on H_0, so a value of $z = 2.182$ would be equally extreme in the upper tail, thus values of z such that $z \geq 2.182$ would *also* be 'at least as extreme as' the observed value $z = -2.182$.

Thus, for this two-sided test, the value of the p-value is

$$\begin{aligned} p &= P(Z \leq -2.182) + P(Z \geq 2.182) \\ &= P(Z \geq 2.182) + P(Z \geq 2.182) \\ &= 2(1 - \Phi(2.182)) \simeq 2(1 - \Phi(2.18)) \\ &= 2(1 - 0.9854) = 2 \times 0.0146 = 0.0292. \end{aligned}$$

The p-value is illustrated in Figure 8 (overleaf).

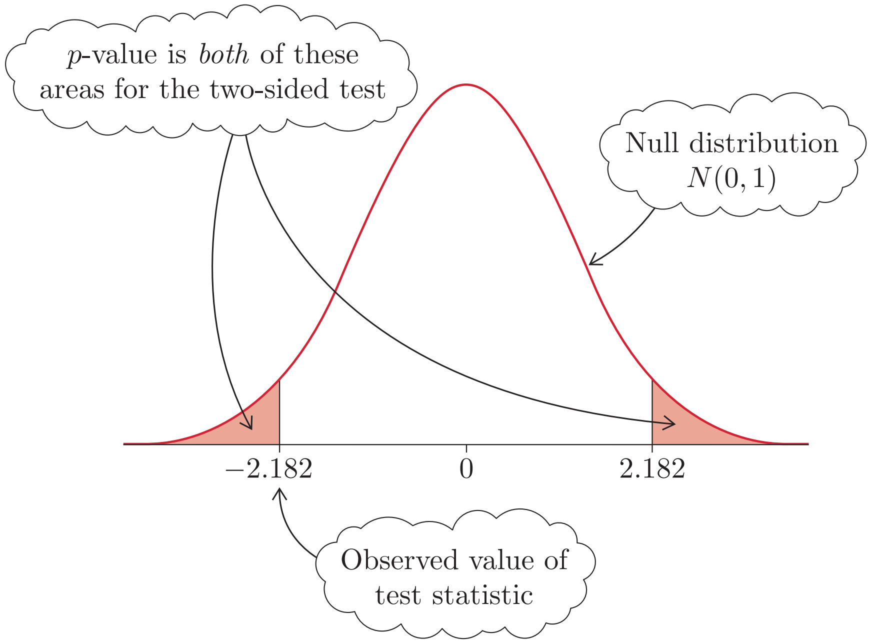


Figure 8 Null distribution $N(0,1)$ with the p-value for the two-sided test for the observed value $z = -2.182$ marked

The value of p is such that $0.01 < p < 0.05$, so from Table 3, there is moderate evidence against H_0. (Note that in Activity 9(a) we rejected H_0 at the 5% significance level.) Further, the fact that $z = -2.182$ suggests that over the period April 2014–March 2015, for females the mean theory test pass rate nationally was lower than the national pass rate for females over the same period the previous year.

Activity 13 *Festive weight gain: p-value*

Activity 7 tested the hypotheses

$$H_0 : \mu = 0.55, \quad H_1 : \mu \neq 0.55,$$

where μ is the mean festive weight gain (in kg). The observed value of the test statistic was $z \simeq -1.657$, and the null distribution was $N(0,1)$.

Calculate and interpret the p-value for this test, and state your conclusions in non-technical language.

In the preceding examples and activities illustrating p-values, the null distribution was $N(0,1)$, so it was easy for us to calculate p-values using normal tables. In general, the null distribution need not be $N(0,1)$ – for example, the null distribution for the test statistic T in Example 6 is $t(19)$. However, the principle for calculating p-values remains the same whatever the null distribution is: the only difference is that a computer is often required to calculate p as tables of sufficient detail are not readily available.

Example 12 *Blueberries and systolic blood pressure: p-value*

Example 6 tested the hypotheses

$$H_0: \mu_S = 138 \text{ mm Hg}, \quad H_1: \mu_S < 138 \text{ mm Hg},$$

where μ_S is the mean systolic BP for menopausal women after taking 22 g of freeze-dried blueberry powder every day for 8 weeks. The observed value of the test statistic for this test was $t \simeq -1.841$, and the null distribution was $t(19)$.

This is a one-sided test, so those values of t which are 'at least as extreme as' its observed value are such that $t \leq -1.841$. Thus

$$p = P(T \leq -1.841),$$

where $T \sim t(19)$. From Minitab, the value of p is calculated to be $p = 0.041$. (You will see how to use Minitab to find this p-value soon.)

The value of p is such that $0.01 < p < 0.05$, so from Table 3, there is moderate evidence against H_0. (Note that in Example 6 we rejected H_0 at the 5% significance level.) We conclude that there is moderate evidence that taking 22 g of freeze-dried blueberry powder every day for 8 weeks lowers systolic BP for menopausal women.

Activity 14 *Two-sided test p-value*

In Example 12, the p-value for the one-sided test of

$$H_0: \mu_S = 138 \text{ mm Hg}, \quad H_1: \mu_S < 138 \text{ mm Hg},$$

was $p = 0.041$.

If instead we wanted to test the hypotheses

$$H_0: \mu_S = 138 \text{ mm Hg}, \quad H_1: \mu_S \neq 138 \text{ mm Hg},$$

what would the p-value be? What do you conclude this time?

In the solution to Activity 14, explicit use was made of something you might have noticed earlier: the p-value associated with a two-sided test is twice the p-value for a corresponding one-sided test.

The general argument which follows applies to both z-tests and t-tests but, for clarity, we will develop it for z-tests only. So, continue to let Z be the random variable version of the test statistic whose observed value is z. The key property of Z is that its null distribution is symmetric about 0.

Consider first a one-sided test of hypotheses of the form

$$H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0,$$

where μ is the population mean and μ_0 is its hypothesised value, and suppose that $z \geq 0$. Then the p-value associated with this test, p_1 say, is

$$p_1 = P(Z \geq z).$$

Consider next a two-sided test of hypotheses of the form

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Then the p-value associated with this test, p_2 say, is the probability that Z is greater than or equal to z (that is, as or more extreme than z in the upper tail) plus the probability that Z is less than or equal to $-z$ (that is, as or more extreme than z in the lower tail). It is because of the symmetry of the null distribution that it is natural to define $-z$ to be as extreme as z in the lower tail, and hence values smaller than $-z$ to be as extreme as values larger than z. In symbols,

$$p_2 = P(Z \geq z) + P(Z \leq -z).$$

However, again because of the symmetry of the null distribution,

$$P(Z \leq -z) = P(Z \geq z);$$

see Figure 9. It follows that

$$p_2 = 2P(Z \geq z) = 2p_1.$$

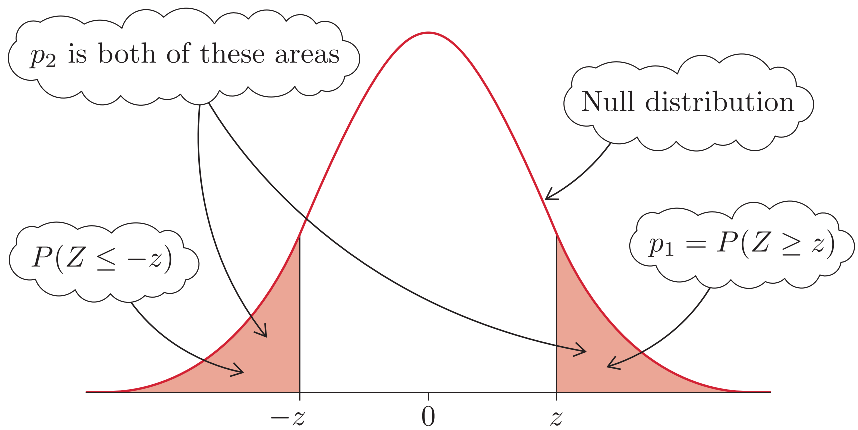


Figure 9 Null distribution with the p-values for the one-sided test, p_1, and for the corresponding two-sided test, p_2, marked

A similar argument applies when the one-sided test is of hypotheses of the form

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0,$$

and $z \leq 0$, but details are omitted.

The main steps in obtaining and using p-values to test hypotheses are summarised in the following box.

The main steps in using p-values for testing hypotheses

- Set up the null and alternative hypotheses.
- Obtain some sample data and summarise these in the test statistic.
- Obtain the null distribution of the test statistic.
- Identify all other values of the test statistic that are at least as extreme, in relation to the null and alternative hypotheses, as the value that was observed.
- Using the null distribution, calculate the p-value as the probability of observing a value of the test statistic at least as extreme as the value observed.
- Interpret the p-value.
- State the conclusion of the test in non-technical language.

4.1 Performing tests and calculating p-values using Minitab

You will now explore how to use Minitab to carry out some standard tests and calculate p-values by working through Chapter 7 of Computer Book B.

Refer to Chapter 7 of Computer Book B for the work in this subsection.

4.2 The link between p-values and rejection decisions

Links, links, ...

Although this section has presented the calculation of significance probabilities or p-values as a different approach to testing hypotheses, it may have become clear to you during the course of the section that there is a strong link between the p-value approach on the one hand, and the rejection decisions made in the approach to testing hypotheses described in Sections 2 and 3 on the other. For instance, in Example 9, the p-value for the test of

$$H_0 : \mu = 51.6\%, \quad H_1 : \mu < 51.6\%,$$

where μ is the mean driving theory test pass rate nationally over the period April 2014–March 2015, was shown to be 0.0455. When interpreting this p-value in Example 10, it was also mentioned that in Example 5 we rejected H_0 at the 5% significance level. How are these two outcomes related?

Let us concentrate on one-tailed tests, like the one just mentioned, with an alternative hypothesis specifying that the parameter of interest is less than

some hypothesised value. Then if Z is the test statistic with observed value z, the p-value is

$$p = P(Z \leq z).$$

On the other hand, the rejection region for a test of significance level $100\alpha\%$ consists of all those values of Z which are less than or equal to c_1, where c_1 is chosen so that

$$\alpha = P(Z \leq c_1).$$

So if z happens to equal c_1, then $p = \alpha$. That is, if $p = \alpha$, you would reject H_0 if you had performed a test at the $100\alpha\%$ significance level.

What if $p < \alpha$? Well, it must then be the case that $z < c_1$. To see this mathematically, $z < c_1$ corresponds to

$$\alpha = P(Z \leq c_1) = P(Z \leq z) + P(z < Z \leq c_1) = p + P(z < Z \leq c_1).$$

So $z < c_1$ corresponds to $p < \alpha$ because $P(z < Z \leq c_1) > 0$. It is probably easier to see this pictorially: the relevant probabilities are shown in Figure 10.

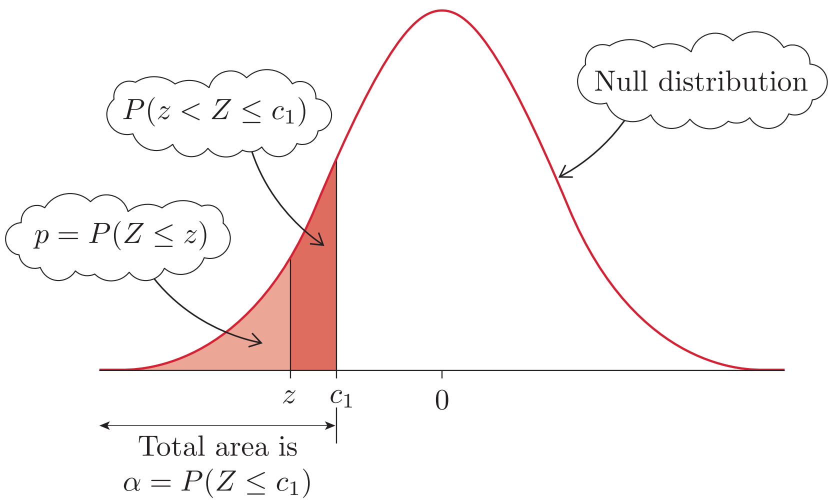


Figure 10 Null distribution with the p-value and rejection region marked when $p < \alpha$

So if $p < \alpha$, it must be the case that z, being less than the critical value c_1, is in the rejection region of an $\alpha\%$ level test, so you would reject H_0 if you had performed a test at the $\alpha\%$ significance level. Combining this with the result for $p = \alpha$ means that if $p \leq \alpha$, you would reject H_0 if you had performed a test at the $100\alpha\%$ significance level.

On the other hand, if $p > \alpha$, you would not reject H_0 if you had performed a test at the $100\alpha\%$ significance level. This is because $p > \alpha$ corresponds to $z > c_1$. Mathematically, we have

$$p = P(Z \leq z) = P(Z \leq c_1) + P(c_1 < Z \leq z) = \alpha + P(c_1 < Z \leq z);$$

pictorially, see Figure 11.

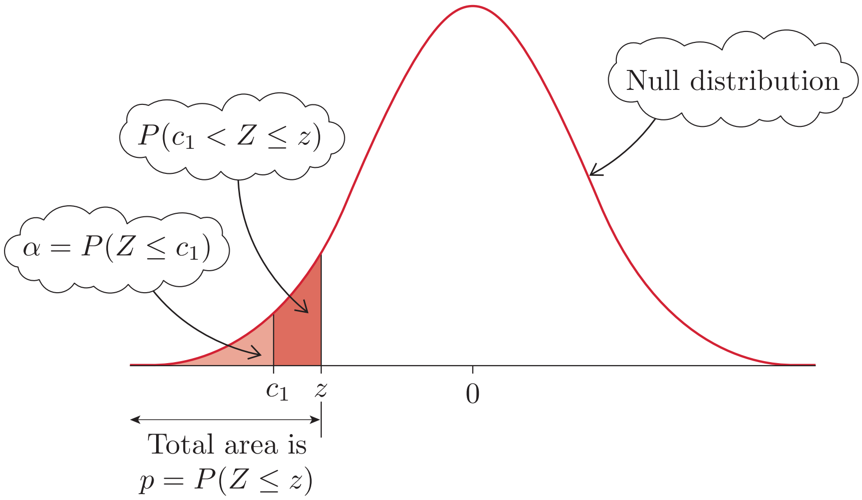


Figure 11 Null distribution with the p-value and rejection region marked when $p > \alpha$

Bringing this all together, it turns out that if you know the p-value, then you know the outcome of a reject/do no reject test at every possible significance level, in the way given in the following box. (Entirely analogous arguments go through for two-sided tests and for one-sided tests with a 'greater than' specification of the alternative hypothesis.)

p-values and rejection decisions

Suppose that a hypothesis test results in a p-value p.

Then a hypothesis test at significance level $100\alpha\%$ would result in:

- the null hypothesis being rejected if $p \leq \alpha$
- the null hypothesis not being rejected if $p > \alpha$.

Notice that the smaller the p-value, the more evidence the data have given us against the null hypothesis, so tests with smaller and smaller significance levels result in rejection of H_0.

The above relationships also make it clear that it is easier to reject a null hypothesis using a test with a higher significance level (e.g. 10%) than it is using a test with a lower significance level (e.g. 1%). This is because the lower the significance level, the more extreme the observed test statistic has to be to be more extreme than the corresponding critical value.

Oh no, it's been rejected; I knew that p-value was too low!

Example 13 *Driving theory test pass rates for females: p-value and rejection decisions*

In Example 11, we reconsidered the two-sided test with hypotheses

$$H_0 : \mu_F = 54.7\%, \quad H_1 : \mu_F \neq 54.7\%,$$

where μ_F is the average driving theory test pass rate for females nationally

over the period April 2014–March 2015. There, using data from Activity 9(a), we calculated that the p-value associated with this test was $p = 0.0292$. As $0.01 < p < 0.05$, this gives moderate evidence against H_0.

We also noted that in Activity 9(a), we had rejected H_0 at the 5% significance level. This accords with the result above: $p = 0.0292 < 0.05 = \alpha$, so knowing that $p = 0.0292$ automatically implies that H_0 would be rejected at the 5% significance level (without having to do the 5% level test). Additionally, we also know that we would not have rejected H_0 if we had performed a 1% level test. This is because $p = 0.0292 > 0.01 = \alpha$ in this case.

In fact, having a p-value of 0.0292 tells us that we would have rejected H_0 if we had performed a hypothesis test at any significance level $100\alpha\%$ such that $\alpha \geq 0.0292$, and that we would not have rejected H_0 if we had performed a hypothesis test at any significance level $100\alpha\%$ such that $\alpha < 0.0292$.

Activity 15 *Blueberries and systolic blood pressure: p-value and rejection decisions*

In Activity 14, you tested the hypotheses

$$H_0 : \mu_S = 138 \text{ mm Hg}, \quad H_1 : \mu_S \neq 138 \text{ mm Hg},$$

where μ_S is the mean systolic BP for menopausal women after taking 22 g of freeze-dried blueberry powder every day for 8 weeks. You found that the p-value associated with this test was $p = 0.082$.

Would you have rejected H_0 or not rejected H_0, if you had tested these hypotheses using each of 1%, 5% and 10% significance levels?

Activity 16 *Interpreting p-values and rejection decisions*

(a) In Table 3, a p-value such that $0.01 < p \leq 0.05$ is interpreted as yielding moderate evidence against a null hypothesis H_0. If a hypothesis test results in moderate evidence against H_0, what can you say about what would have happened had you tested these hypotheses using each of 1%, 5% and 10% significance levels?

(b) Repeat part (a) for a test whose p-value provides strong evidence against H_0.

Suppose now that you performed a hypothesis test at, say, the 5% level and decided to reject H_0. Was this a marginal decision or a very clear-cut one? Would you have also rejected the test at the 1% level or not? Another advantage of the p-value approach to hypothesis testing is that the answer to these questions is clearly given by the p-value. To illustrate, if $p = 0.049$, then you know that the decision to reject H_0 at the 5% level

is marginal and that you would not have also rejected the test at the 1% level (or indeed at any significance level less than or equal to 4.9%). For example, consider again the festive weight gain example of Activity 8; there, the null hypothesis of a weight gain of 0.55 kg was rejected against a weight gain of less than 0.55 kg at the 5% significance level because, from the solution to Activity 8, 'the observed value of the test statistic is $z = -1.657$ and $-1.657 < -1.645$ (just), [so] the observed value of the test statistic is in the rejection region'. In fact, the p-value in this case is 0.049, confirming that the decision to reject was, in that case, a marginal one.

Indeed, rejection decisions, unless qualified as marginal, can mislead. For example, p-values of 0.050001 and 0.050000 are essentially the same but would result in different decisions in a test at the 5% significance level, the first not to reject H_0, the second to reject H_0. Giving the p-value makes the true situation much clearer.

Exercises on Section 4

Exercise 5 *Insect traps: p-value*

Exercise 2 considered data on the number of insect traps containing insects of the taxon *Staphylinoidea*. A z-test was carried out to test the hypotheses

$$H_0 : \mu = 1, \quad H_1 : \mu > 1,$$

where μ is the underlying mean number of insects of the taxon *Staphylinoidea* in the traps.

The observed value of the test statistic was $z \simeq 2.208$. Calculate the associated p-value for the test, and interpret its value.

Exercise 6 *Spelling skills of braille readers*

Exercise 3 considered data on the spelling skills of visually impaired children who read using braille. A normal model was assumed for the variation in the test scores of this small sample of data. A t-test was carried out of the hypotheses

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100,$$

where μ is the mean spelling test score of visually impaired children using braille, using data from a sample of 23 such children. The observed value of the test statistic was $t \simeq -0.410$.

(a) Obtain an expression for the p-value associated with this test, giving the distribution under which any probabilities are to be calculated. (Do not attempt to evaluate any such probabilities.)

(b) The numerical value of the p-value obtained from the expression that you should have obtained in part (a) is 0.6858. Interpret this p-value.

Exercise 7 *Interpreting weak evidence and rejection decisions*

In Table 3, a p-value such that $0.05 < p \leq 0.10$ is interpreted as yielding weak evidence against a null hypothesis H_0. If a hypothesis test results in weak evidence against H_0, what can you say about what would have happened had you tested these hypotheses using each of 1%, 5% and 10% significance levels?

5 Power, and choosing the sample size

Despite all the positive things said about the p-value approach to hypothesis testing in Section 4, in this section we will return to the context of tests based on significance levels and rejection regions, that is, using the approach to testing hypotheses presented in Sections 2 and 3. In such tests, there are two possible decisions: either we reject H_0, or we do not reject H_0. However, it is possible that we make the incorrect decision. We begin this section by considering the two possible errors that can occur when rejecting, or not rejecting, H_0.

5.1 Type I and Type II errors

Muscle fibres come in Types I and II too. Type I muscles are 'slow twitch'; Type II muscles are 'fast twitch'. The athletes illustrated here have particularly honed their Type I muscles which are responsible for endurance.

When testing hypotheses, there are two errors that can be made:

- Reject H_0 when H_0 is true – this is called a **Type I error**
- Do not reject H_0 when H_0 is false – this is called a **Type II error**.

The names given to the types of error are, unfortunately, not very descriptive or easy to remember. It might help to think of Type I errors as corresponding to *false positives*: by rejecting the null hypothesis, you think you have evidence of something non-null – a 'positive' result – but if the test is in error, the result is a false positive. Similarly, Type II errors correspond to *false negatives*: the hypothesis test provides little or no evidence against the null hypothesis – a 'negative' result – but if the test is in error, the result is a false negative.

When carrying out a test, we cannot know whether we have made either of these errors because we do not know whether H_0 is true or not (if we did know, then there would be no point in performing the test!). We can, however, consider the probability of making a Type I error and the probability of making a Type II error.

So, recall that H_0 is rejected if the test statistic lies in the rejection region. The rejection region is calculated by setting the significance level, $100\alpha\%$, to a given value, usually 1%, 5% or 10%. Equivalently, the probability α is set to 0.01, 0.05 or 0.1. Notice that the significance level might be specified by either a percentage ($100\alpha\%$) or a probability (α). From here on, it makes formulas easier to use the phrase 'significance level' to mean its level

on the probability scale rather than the percentage scale. The null distribution – that is, the distribution of the test statistic when H_0 is true – is then used to calculate the rejection region so that

$$P(\text{test statistic lies in rejection region}) = \alpha = \text{significance level}.$$

This is illustrated for a two-sided test in Figure 12.

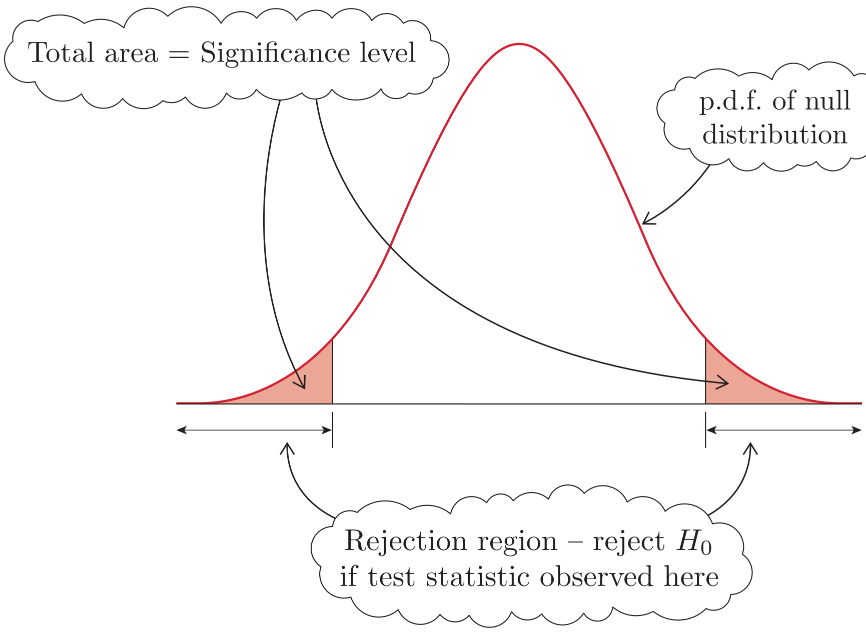


Figure 12 The p.d.f of the null distribution with the rejection region and significance level marked

But if the test statistic lies in the rejection region, then we reject H_0. Continuing to use 'significance level' on the probability rather than percentage scale, the significance level is therefore the probability of rejecting H_0 when H_0 is true, so

$$\text{significance level} = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P(\text{Type I error}).$$

The significance level is chosen by the person carrying out the test. The probability of a Type I error is therefore within the control of the designer of the test. It is clearly desirable to choose this probability to be small.

Now consider the probability of a Type II error – that is,

$$P(\text{Type II error}) = P(\text{do not reject } H_0 \text{ when } H_0 \text{ is false}).$$

This probability is also determined by the rejection region, which in turn is controlled by the person carrying out the test through the significance level.

The good news is that the smaller the rejection region, the less likely it is that a Type I error will occur. But the bad news is that, for fixed sample size n, the smaller the rejection region, the more likely it will be that a Type II error could be made (that is, H_0 isn't rejected when in fact it should be). There is thus a trade-off to be made between these two error probabilities when designing a test.

You will see this for yourself in the next subsection.

The two types of error are summarised in the following box.

Possible errors when testing hypotheses

- A **Type I error** occurs when we reject H_0 but it is true. It is the case that

  $$\text{significance level} = P(\text{Type I error}).$$

- A **Type II error** occurs when we do not reject H_0 but it is false.
- There is a trade-off between the two error probabilities when designing a test.

Refer to Chapter 8 of Computer Book B for the rest of the work in this subsection.

5.2 The power of a test

In many circumstances, the usefulness of a test is measured by what is called the **power** of the test. This is defined as the probability of making the correct decision when the null hypothesis is not true, which in this case is to reject the null hypothesis:

$$\begin{aligned} \text{power} &= P(\text{reject } H_0 \text{ when } H_0 \text{ is false}) \\ &= 1 - P(\text{do not reject } H_0 \text{ when } H_0 \text{ is false}) \\ &= 1 - P(\text{Type II error}). \end{aligned}$$

The athletes illustrated here have particularly honed their Type II muscles which are responsible for ... power. They seem to be avoiding Type II errors!

Thus the power is also the probability of *avoiding* a Type II error.

Clearly, we would like the power to be as large as possible because it measures the effectiveness of the test when the null hypothesis is not true.

The power of a test

- The **power** of a test is

  $$\text{power} = P(\text{reject } H_0 \text{ when } H_0 \text{ is false}),$$

  which is also the probability of avoiding a Type II error:

  $$\text{power} = 1 - P(\text{Type II error}).$$

- It is desirable to have large power and small significance level.

Unfortunately, the power is directly related to the probability of a Type II error, and we noted in Subsection 5.1 that there is a trade-off between the probability of a Type I error (that is, the significance level) and the probability of a Type II error. Hence there is also a trade-off between the significance level and the power: the larger the power (which is desirable), the larger the probability of a Type I error (which is undesirable).

To see this another way, notice that both the significance level and the power are probabilities of rejecting H_0. They differ because the

significance level is the probability of rejecting H_0 when H_0 is true, and the power is the probability of rejecting H_0 when H_0 is false. An extreme example of the inability to make both significance level small and power large is considered in the following activity.

Activity 17 *Always reject!*

Suppose that someone suggested that they are going to make the decision 'reject H_0' regardless of what the data or any resulting test statistic tells them. Such a test is clearly unacceptable on intuitive grounds! Explain why the test is also unacceptable on technical grounds, in terms of the power and/or the significance level of such a test.

There is, however, one way to simultaneously increase the power and reduce the probability of a Type I error, and that is by increasing the sample size. In situations in which the sample size is under the control of the person gathering the data, calculations involving the power can throw light on what an appropriate sample size would be. On the one hand, it makes very little sense to spend time and money on gathering data that are unlikely to lead to the null hypothesis being rejected even when it is false, that is, on performing a test that has low power. On the other hand, if a planned study has extremely high power, it might be a better use of resources to reduce the sample size and spend the resulting savings on something else. Thus institutions responsible for approving and funding research often require to see appropriate statistical power calculations as part of any research plan that they consider.

5.3 Calculating power

Some of the details of power calculations can be complicated and are best left to a computer. Therefore, only one particular test situation is considered in this section, while further situations are considered in Computer Book B. Even in this situation, what matters most is not so much that you are able to follow every mathematical detail in the development, but that you can use the derived formulas for making power calculations.

The eagle is the world's most powerful bird: it can carry up to four times its own weight while flying

Suppose that a sample of size n can be modelled by a normal distribution $N(\mu, \sigma^2)$, where the standard deviation, σ, is known. In this case, the sample mean $\overline{X}$ follows the $N(\mu, \sigma^2/n)$ distribution. Suppose further that we wish to test the hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0,$$

using the significance level α and the test statistic

$$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Notice that this test statistic is very similar to the Z and T test statistics in the z- and t-tests of Subsection 3.1. The only difference is that the standard deviation was estimated by s in Z and T, whereas it is assumed known here.

Now, when H_0 is true so that $\mu = \mu_0$,

$$\overline{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right),$$

Notice that this result holds for *any* sample size n, since the data are normally distributed and σ is assumed known.

so

$$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

With significance level α, the critical value for the test is then the value c such that

$$P(Z_1 \geq c \text{ when } H_0 \text{ is true}) = \alpha.$$

So, since $Z_1 \sim N(0,1)$ when H_0 is true, the critical value c is $q_{1-\alpha}$, the $(1-\alpha)$-quantile of $N(0,1)$. We thus reject H_0 if the observed value of the test statistic z_1 is such that $z_1 \geq q_{1-\alpha}$.

This is illustrated in Figure 13.

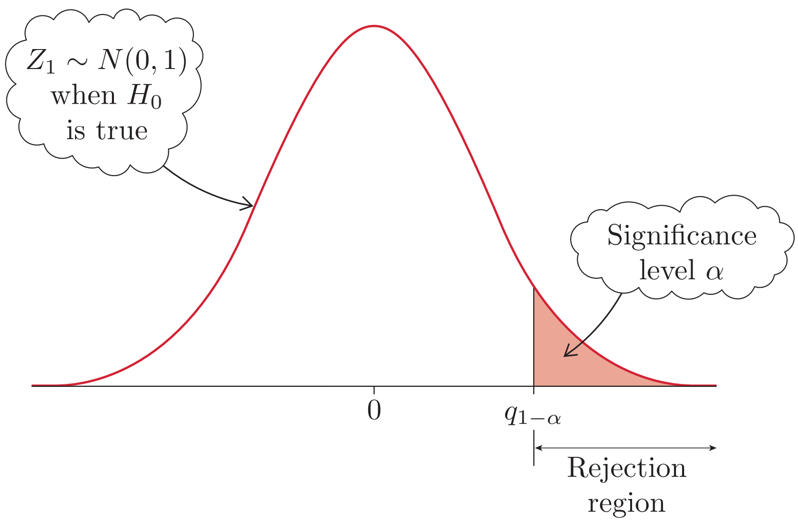


Figure 13 Critical value and rejection region when using test statistic Z_1

Notice that, for the moment, d is a fixed value that needs to be specified in order to calculate the power.

Suppose now that μ is in fact $\mu_0 + d$, for some $d > 0$, so that μ is one of the values under H_1, and H_0 is actually false. In this case, the power of the test is

$$\begin{aligned} \text{power} &= P(\text{reject } H_0 \text{ when } \mu = \mu_0 + d) \\ &= P(Z_1 \geq q_{1-\alpha} \text{ when } \mu = \mu_0 + d). \end{aligned}$$

Although the power and the significance level both use $P(Z_1 \geq q_{1-\alpha})$, they are *not* the same probability. When considering the significance level, the distribution of Z_1 is $N(0,1)$: this is the distribution of Z_1 when H_0 is true. (And so in this case $P(Z_1 \geq q_{1-\alpha}) = \alpha$.)

However, when considering the power, $\mu = \mu_0 + d$, and the distribution of Z_1 is *not* $N(0,1)$. So what is the distribution of Z_1 when $\mu = \mu_0 + d$? If $\mu = \mu_0 + d$, then

$$\overline{X} \sim N\left(\mu_0 + d, \frac{\sigma^2}{n}\right),$$

so that, standardising as usual by subtracting the mean and dividing by the standard deviation,

$$\frac{\overline{X} - (\mu_0 + d)}{\sigma/\sqrt{n}} \sim N(0,1).$$

But

$$\frac{\overline{X} - (\mu_0 + d)}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} - \frac{d}{\sigma/\sqrt{n}} = Z_1 - \frac{d}{\sigma/\sqrt{n}},$$

so when $\mu = \mu_0 + d$,

$$Z_1 - \frac{d}{\sigma/\sqrt{n}} \sim N(0,1).$$

Write

$$W = Z_1 - \frac{d}{\sigma/\sqrt{n}}$$

so that, equivalently, $W \sim N(0,1)$.

Although it will still be useful to work out the distribution of Z_1 shortly, we don't actually need it explicitly to work out the power after all; we can use the (simpler) distribution of W instead:

$$\begin{aligned}
\text{power} &= P(Z_1 \geq q_{1-\alpha} \text{ when } \mu = \mu_0 + d) \\
&= P\left(Z_1 - \frac{d}{\sigma/\sqrt{n}} \geq q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right) \\
&= P\left(W \geq q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right) \\
&= 1 - P\left(W < q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right) \qquad (1)
\end{aligned}$$

since $W \sim N(0,1)$. So Equation (1) is the formula that we were seeking for the power of the test of the hypotheses

$$H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0;$$

when the data are from the $N(\mu, \sigma^2)$ distribution, we use the significance level α, and $\mu = \mu_0 + d$.

Figure 14 (overleaf) illustrates this power. Notice that the significance level relates to the distribution of Z_1 when H_0 is true so that $\mu = \mu_0$, while the power relates to the distribution of Z_1 when $\mu = \mu_0 + d$. The latter distribution is named on Figure 14; you will show that this is the correct distribution in the following activity.

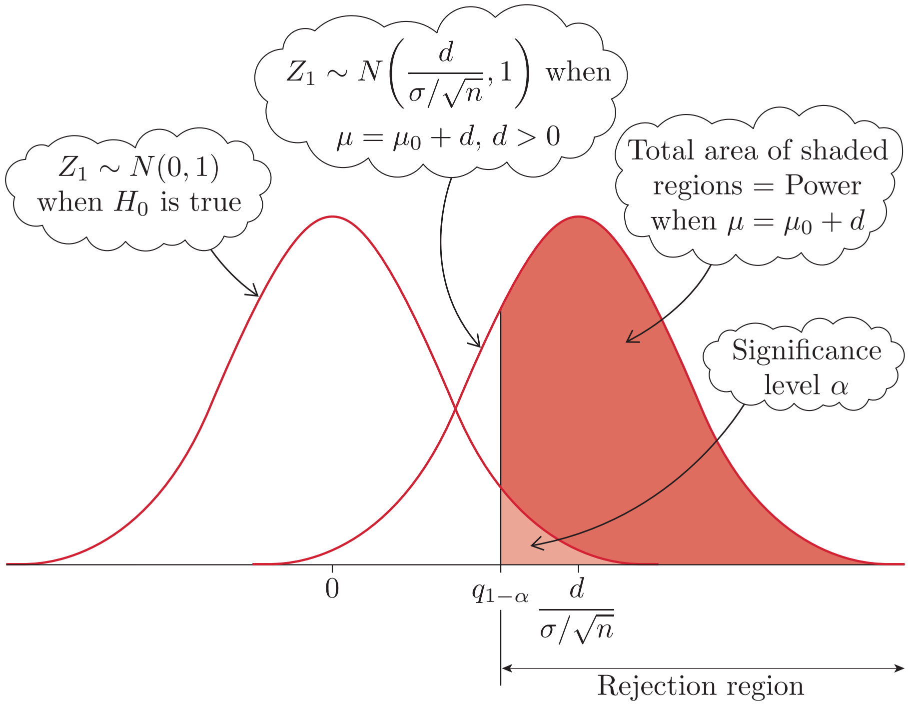


Figure 14 Illustration of the power when $\mu = \mu_0 + d$

Activity 18 *Distribution of Z_1*

From Subsection 3.1 of Unit 6, if $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$, for any constants a, b. Use this result and the relationship between Z_1 and W to show that the distribution of Z_1 when $\mu = \mu_0 + d$ is

$$Z_1 \sim N\left(\frac{d}{\sigma/\sqrt{n}}, 1\right).$$

Calculating a power is illustrated in Example 14.

The gorilla is the most powerful large animal in the world: it can lift up to ten times its own body weight

Example 14 *A power calculation*

Suppose that a sample of size $n = 25$ is obtained from a population distributed as $N(\mu, 100)$ and we wish to test the hypotheses

$$H_0 : \mu = 5, \quad H_1 : \mu > 5.$$

Suppose that the actual mean of the population is 7. Then since $\mu_0 = 5$ and $\mu_0 + d = 7$, the value of d is 2 and

$$\frac{d}{\sigma/\sqrt{n}} = \frac{2}{10/\sqrt{25}} = 1.$$

Therefore, when the true value of μ is 7, $Z_1 \sim N(1,1)$ and, from Equation (1), the power for significance level 0.05 is calculated as

$$\begin{aligned}\text{power} &= 1 - \Phi\left(q_{0.95} - \frac{d}{\sigma/\sqrt{n}}\right)\\ &= 1 - \Phi(1.645 - 1) = 1 - \Phi(0.645) \simeq 1 - \Phi(0.65).\end{aligned}$$

From standard normal tables, $\Phi(0.65) = 0.7422$, so the power is approximately $1 - 0.7422 = 0.2578$. The probability of correctly rejecting H_0 when μ is 7 is therefore not very high!

The power for this test is illustrated in Figure 15.

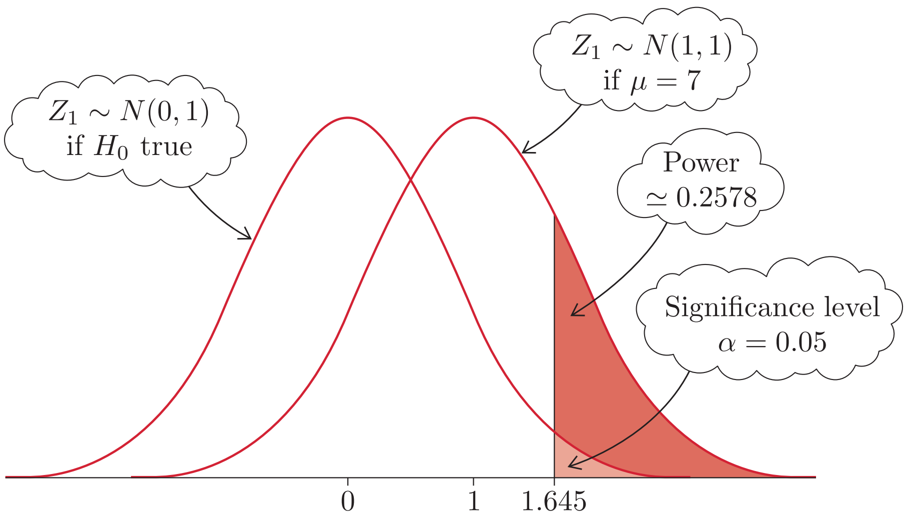


Figure 15 Illustration of the power when testing hypotheses $H_0 : \mu = 5$, $H_1 : \mu > 5$, when the true value of μ is 7

The explicit calculations that have been made in this subsection so far allow us to take another look at the trade-off between significance level and power. In the following screencast, you will see how changing the probability of the Type I error (which is in the control of the person conducting the test) affects the power of the test in the current context.

Screencast 9.2 The trade-off between significance level and power

So far, we have considered only the alternative hypothesis $H_1 : \mu > \mu_0$ with the true value of μ being greater than μ_0 (that is, $\mu = \mu_0 + d$, for $d > 0$). In the next activity, you will obtain the power in the case of $H_1 : \mu < \mu_0$ and the true value of μ is $\mu_0 - d$, for $d > 0$.

Activity 19 *Calculating power for alternative hypothesis $H_1: \mu < \mu_0$*

A sample of size n is obtained from a population distributed as $N(\mu, \sigma^2)$, where σ^2 is assumed known. Suppose that we wish to test the null hypothesis $H_0: \mu = \mu_0$ using test statistic $Z_1 = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ and significance level α.

Suppose that the alternative hypothesis is $H_1: \mu < \mu_0$ and the true value of μ is $\mu = \mu_0 - d$, where $d > 0$.

(a) What would the rejection region be for this test?

(b) Define

$$V = Z_1 + \frac{d}{\sigma/\sqrt{n}}.$$

Show that when $\mu = \mu_0 - d$, where $d > 0$, the distribution of V is $N(0,1)$.

(c) Hence show that the power when $\mu = \mu_0 - d$, where $d > 0$, is still calculated as

$$\text{power} = 1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right).$$

Now suppose that the alternative hypothesis is $H_1: \mu \neq \mu_0$ so that we have a two-sided test. In this case, the rejection region will be values of z_1 such that $z_1 \leq -q_{1-(\alpha/2)}$ and $z_1 \geq q_{1-(\alpha/2)}$.

Suppose that the true value of μ is $\mu_0 + d$, where $d > 0$. In this case, $Z_1 \sim N(\frac{d}{\sigma/\sqrt{n}}, 1)$ (as in Activity 18) and, in particular, unless d is small, $P(Z_1 \leq -q_{1-(\alpha/2)})$ will be effectively zero. Thus, for a two-sided test where the true value of μ is $\mu_0 + d$, where $d > 0$ is not small,

$$\begin{aligned} \text{power} &\simeq P(Z_1 \geq q_{1-(\alpha/2)} \text{ when } \mu = \mu_0 + d) \\ &= 1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right). \end{aligned} \tag{2}$$

(Equation (2) is a consequence of the same argument that led to Equation (1) except that $q_{1-\alpha}$ there is replaced by $q_{1-(\alpha/2)}$ here.) This is illustrated in Figure 16.

If the true value of μ is instead $\mu_0 - d$, where $d > 0$, then, this time, it can be shown by a similar argument to that in Activity 18 (but which you can take on trust) that $Z_1 \sim N(-\frac{d}{\sigma/\sqrt{n}}, 1)$ and, unless d is small, $P(Z_1 \geq q_{1-(\alpha/2)})$ will be effectively zero. Then, following similar arguments to those of Activity 19 but with $q_{1-\alpha}$ replaced by $q_{1-(\alpha/2)}$, the power is still calculated as

$$\text{power} = 1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right).$$

This is illustrated in Figure 17.

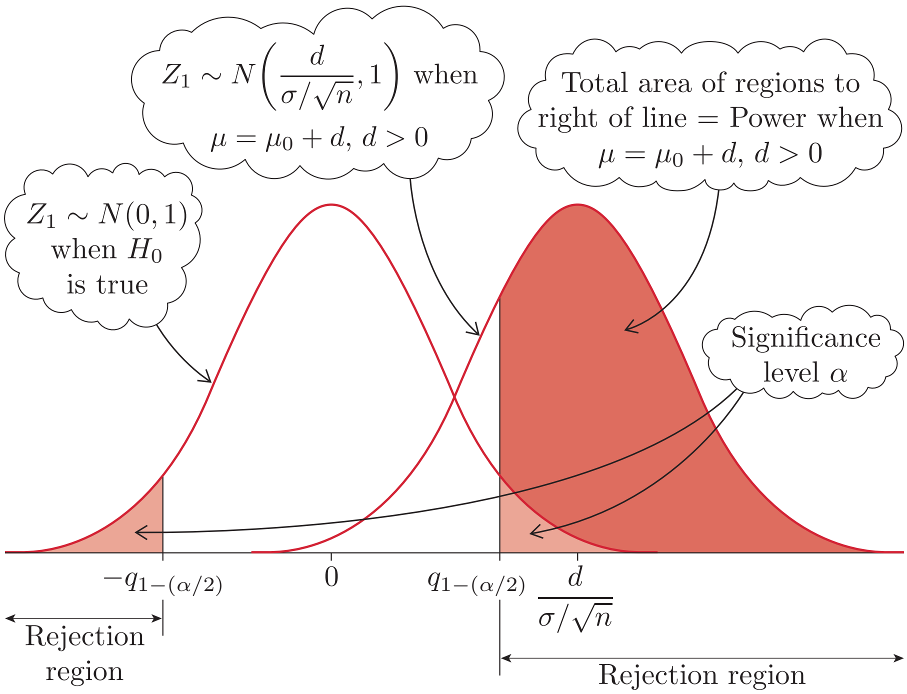


Figure 16 Illustration of the power when testing hypotheses $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, when the true value of μ is $\mu_0 + d$, $d > 0$

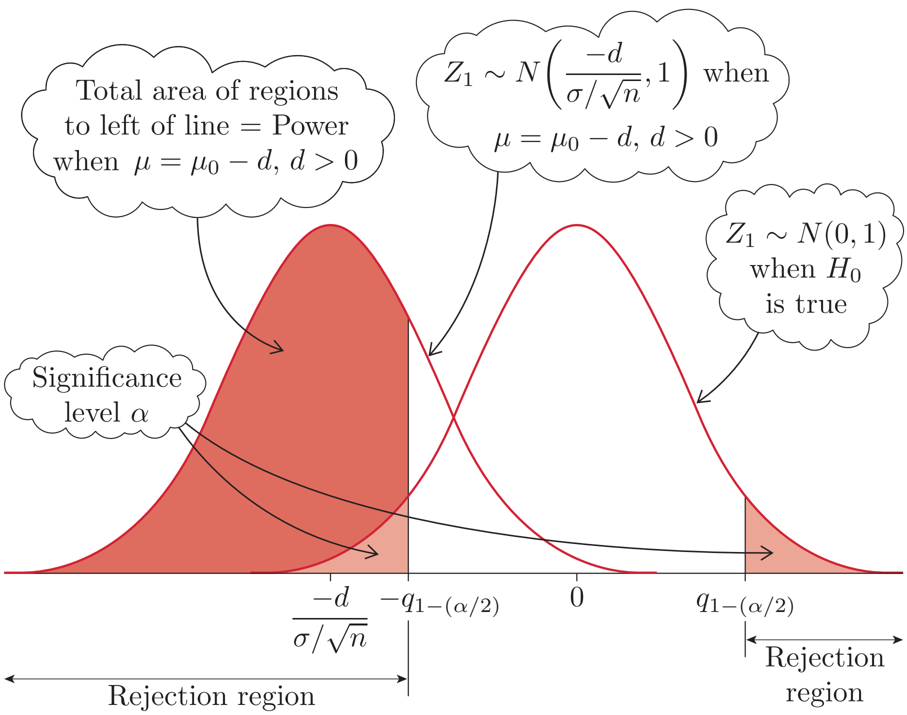


Figure 17 Illustration of the power when testing hypotheses $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, when the true value of μ is $\mu_0 - d$, $d > 0$

The power calculations presented in this subsection are summarised in the following box.

Power calculations

Suppose that a sample of size n is obtained from a population distributed as $N(\mu, \sigma^2)$, where σ^2 is assumed known, and the test statistic

$$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

is to be used in a test of the null hypothesis $H_0 : \mu = \mu_0$ with significance level α. Let $d > 0$.

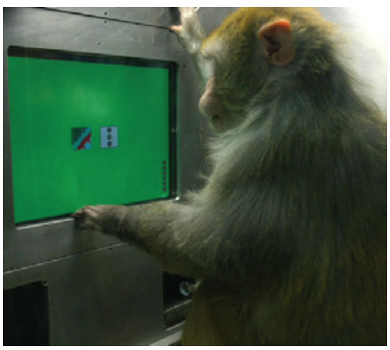
- When the alternative hypothesis is of the form $H_1 : \mu > \mu_0$ and the true value of μ is $\mu_0 + d$, or when the alternative hypothesis is of the form $H_1 : \mu < \mu_0$ and the true value of μ is $\mu_0 - d$, the **power of the one-sided test** is

$$1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right).$$

- When the alternative hypothesis is of the form $H_1 : \mu \neq \mu_0$, the **power of the two-sided test** when the true value of μ is $\mu_0 \pm d$, where d is not small, is approximately

$$1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right).$$

Activity 20 *Testing IQs*



An IQ test for a different population?

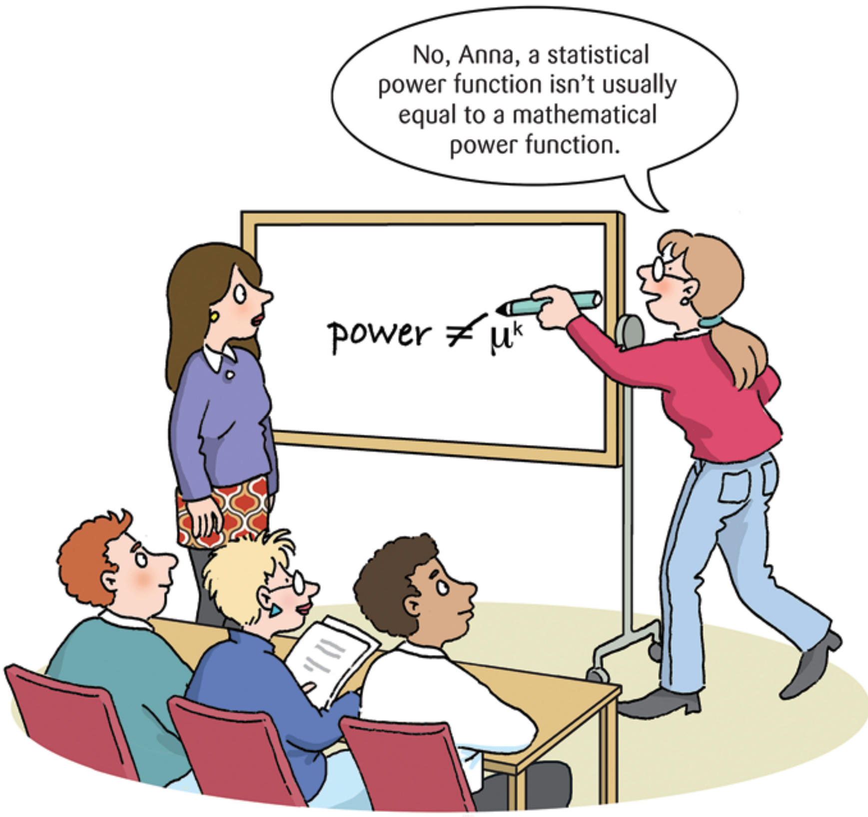
A psychologist wishes to investigate the IQ of a certain specific population. The aim is to investigate whether the mean IQ in this population could plausibly be equal to that in the population of the UK as a whole. For the IQ test that the psychologist will use, the scores for the general UK population are distributed as $N(100, 15^2)$. The psychologist intends to take a sample of 80 individuals from the specific population and measure their IQs using this test. It is thought likely that the standard deviation of IQ scores in the specific population is the same as that in the general UK population (though, of course, nobody can be sure until the data have been collected). The psychologist will test the null hypothesis that the mean IQ in the specific population is 100, using a two-sided test, at significance level 0.05. Suppose that the actual mean IQ score in the specific population is 104.75. What is the probability that the psychologist will reject the null hypothesis?

Equations (1) and (2) for calculating the power are, strictly, valid only if the population standard deviation is known. In practice, in the great majority of cases, it will not be known, and the population standard

deviation will be replaced by the sample standard deviation and a t-test performed. If the sample size is fairly large, these expressions will still give reasonable approximations to the power of the test; but if the sample size is small, they will not. More complicated calculations, based on the same principles, can give accurate values for the power of t-tests; you will use your computer later for such calculations.

The power function

When calculating the power, we calculated the probability of rejecting the null hypothesis for a specified 'true' value of μ. However, in reality we won't know what μ is, so we would like to know what the power is for many possible values of μ. The **power function** is simply a function which gives the power of a test for possible values of μ.



Example 15 *Power function for a one-sided test*

Consider once again the scenario presented in Example 14 in which a sample of size $n = 25$ is obtained from a population distributed as $N(\mu, 100)$ and we wish to test the hypotheses

$$H_0 : \mu = 5, \quad H_1 : \mu > 5,$$

at the 5% significance level.

Figure 18 shows the power function for this test plotted as a function of μ. Its formula is given by Equation (1) with d set to $\mu - 5$ (since $\mu = \mu_0 + d$ and $\mu_0 = 5$), $\sigma = 10$ and $n = 25$. That is,

$$\begin{aligned}\text{power} &= 1 - \Phi\left(q_{0.95} - \frac{d}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(1.645 - \frac{\mu - 5}{10/\sqrt{25}}\right)\\ &= 1 - \Phi\left(4.145 - \frac{\mu}{2}\right).\end{aligned}$$

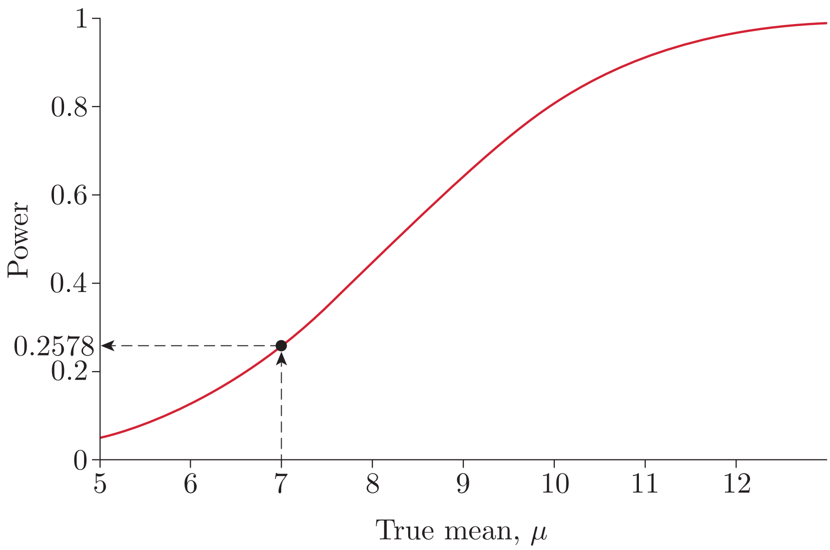


Figure 18 Power function for a one-sided test ($H_0 : \mu_0 = 5$, $H_1 : \mu > 5$, significance level 0.05)

Notice from the figure that the power of the test increases as the true value of μ increases and moves further away from the hypothesised value $\mu = 5$.

Recall that in Example 14 it was assumed that the true mean μ is 7, and in this case the power was 0.2578. Notice that the value of the power function when $\mu = 7$ is indeed 0.2578.

Activity 21 *The lowest power*

The lowest point on the graph of the power function occurs when the true mean is equal to the hypothesised mean (that is, $\mu = 5$). In this case the power is 0.05. Can you explain why this is?

Activity 22 *Power function for a two-sided test*

If a two-sided test is used instead of a one-sided test in the scenario of Example 14 (so that we have $H_1 : \mu \neq 5$), then the power function is as shown in Figure 19.

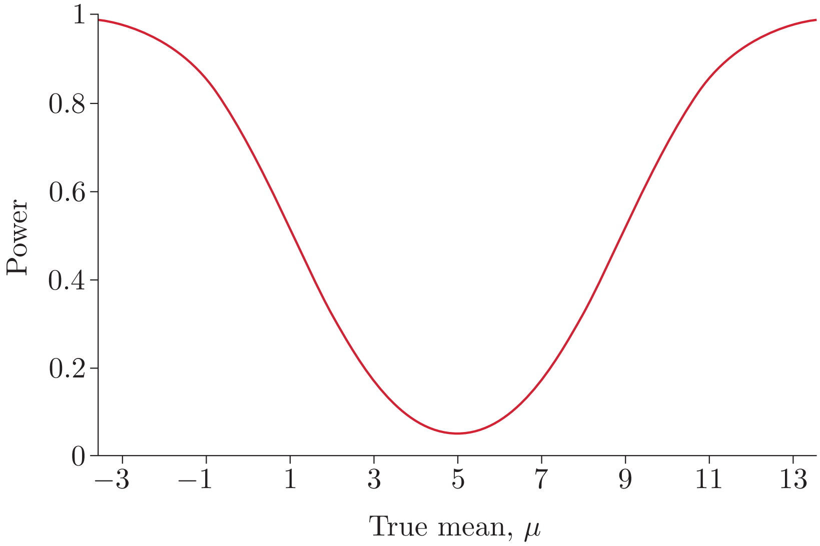


Figure 19 Power function for a two-sided test ($H_0 : \mu = 5$, $H_1 : \mu \neq 5$, significance level 0.05)

Can you explain why the power function for this two-sided test has the shape that it does?

5.4 Planning sample sizes

Perhaps the most common use of power calculations for a hypothesis test is in order to get an idea of what appropriate sample sizes would be for tests to provide useful results efficiently in practice. In the previous subsection, the power of a test was calculated for a sample of data of size n: in this subsection we turn this around and calculate what the sample size needs to be in order for a test to have a particular power.

Taking a red sample!

Once again we'll consider the test scenario in which a sample of size n can be modelled by a normal distribution $N(\mu, \sigma^2)$, where σ^2 is known, and the test statistic

$$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

will be used to test the hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0.$$

In the previous subsection, you saw that in this case, when the true value of μ is $\mu_0 + d$ and the significance level is α,

$$\begin{aligned} \text{power} &= P(Z_1 \geq q_{1-\alpha} \text{ when } \mu = \mu_0 + d) \\ &= 1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right). \end{aligned}$$

This is Equation (1).

Suppose that we wish the power of our test to be some value γ, say; in practice, γ is often taken to be something like 0.8 or 0.9. Then we require that

$$\gamma = 1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right)$$

and hence

$$1 - \gamma = \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right). \tag{3}$$

Now, recall the definition of a standard normal quantile from Unit 6: q_β is the 100β%-quantile of $N(0,1)$ if it satisfies

$$\beta = \Phi(q_\beta).$$

Equation (3) therefore asks that

$$q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}} = q_{1-\gamma},$$

where $q_{1-\gamma}$ is the $(1-\gamma)$-quantile of $N(0,1)$.

We can now rearrange this formula so that we get an expression for n:

$$q_{1-\alpha} - q_{1-\gamma} = \frac{d}{\sigma/\sqrt{n}}$$

so

$$(q_{1-\alpha} - q_{1-\gamma})\frac{\sigma}{d} = \sqrt{n}$$

thus

$$n = \frac{\sigma^2}{d^2}(q_{1-\alpha} - q_{1-\gamma})^2. \tag{4}$$

Example 16 *Sample size calculation*

Consider once again Example 14 in which a sample of size $n = 25$ is obtained from a population distributed as $N(\mu, 100)$. The hypotheses to be tested are

$$H_0: \mu = 5, \quad H_1: \mu > 5,$$

and the actual mean of the population is 7, so $\mu_0 + d = 7$ and the value of d is therefore 2. In that example, when the significance level is $\alpha = 0.05$, the power is only 0.2578.

Now suppose that we wish to calculate the sample size required for the power to be 0.9. We can do this using Equation (4). The significance level is 0.05, so $q_{1-\alpha} = q_{0.95} = 1.645$. We wish the power to be 0.9, so set γ to be 0.9 so that $q_{1-\gamma} = q_{0.1} = -q_{0.9} = -1.282$. Then

$$\begin{aligned} n &= \frac{\sigma^2}{d^2}(q_{1-\alpha} - q_{1-\gamma})^2 = \frac{100}{2^2}(1.645 - (-1.282))^2 \\ &= 25(1.645 + 1.282)^2 \simeq 214.2. \end{aligned}$$

So in order for the power to be 0.9, $n \simeq 214.2$. The sample size needs to be a whole number, so this should be rounded up to 215. (By rounding up

rather than down, we ensure that the power of the test is at least what we want it to be – in this case, 0.9.)

The form of Equation (4) is interesting. The term $(q_{1-\alpha} - q_{1-\gamma})^2$ is quite substantial since, typically, $q_{1-\alpha}$ is quite a high quantile of $N(0,1)$ (α tends to be small) and $q_{1-\gamma}$ is quite a low quantile of $N(0,1)$ (γ tends to be quite large); it turned out to be a bit over 8.5 in Example 16. Equation (4) also tells us how the sample size should be changed in response to changes in both the variability of the underlying population and the difference in means that we wish to be detectable. You can explore this for yourself in the next activity.

Activity 23 *How does sample size depend on variability and difference between means?*

(a) If the amount of variability in the data increases, and everything else stays the same, should you increase or decrease the sample size in order to achieve the same power? More specifically, if you decide that the standard deviation is twice what you originally thought it might be, by what factor should you adjust the required sample size?

(b) If the difference between means that you wish the test to detect decreases, and everything else stays the same, should you increase or decrease the sample size in order to achieve the same power? More specifically, if you decide that you would like to detect a difference in means that is half what you originally desired, by what factor should you adjust the required sample size?

In both parts of the question, ignore the fact that you may round your final sample size up.

In the following activity, you will consider how to calculate the sample size for a given power for alternative hypotheses $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$.

Activity 24 *Sample size calculations for $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$*

(a) If, instead of the situation discussed so far in this subsection, the hypotheses being tested are

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0,$$

confirm that the sample size n for given power γ is still calculated using Equation (4).

(b) If instead the hypotheses being tested are

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

confirm that the sample size n for given power γ is calculated using

$$n = \frac{\sigma^2}{d^2}(q_{1-(\alpha/2)} - q_{1-\gamma})^2.$$

These expressions for calculating the sample size for a given power also give approximately correct answers in the case where the underlying variance is not known and a t-test is being performed. In this case, the approximation is reasonable provided that d/σ is not too small.

The sample size calculations are summarised in the following box.

Choosing the sample size

Suppose that a sample of size n is to be obtained from a population distributed as $N(\mu, \sigma^2)$, where σ^2 is assumed known, and the test statistic

$$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

is to be used in a test of the null hypothesis $H_0 : \mu = \mu_0$ with significance level α. Suppose further that the true underlying mean is $\mu_0 + d$.

The sample size n required so that the power of the test is equal to a predetermined value γ is:

- for a one-sided test,

$$n = \frac{\sigma^2}{d^2}(q_{1-\alpha} - q_{1-\gamma})^2$$

- for a two-sided test,

$$n = \frac{\sigma^2}{d^2}(q_{1-(\alpha/2)} - q_{1-\gamma})^2.$$

Activity 25 *Sample size for a drug investigation*

A researcher plans an investigation of whether a particular drug alters blood pressure. In the investigation, each member of a group of individuals will have their blood pressure measured; then they will take the drug, and one hour later they will have their blood pressure measured again. Suppose that it is known that for patients who have not taken any drug, the standard deviation of hourly systolic blood pressure measurements on the same patient is about 10 mm Hg. The researcher intends to analyse the data by applying a two-sided test (based on the normal distribution) to the individual differences in blood pressure before and after taking the drug, at a significance level of 0.01. The intention is that the study should have a power of 0.9 for finding a mean difference in systolic blood pressure of 5 mm Hg. How many participants should the researcher use?

Refer to Chapter 9 of Computer Book B for the rest of the work in this subsection.

Exercises on Section 5

Exercise 8 *Power*

Suppose that the researcher in Activity 25 did indeed carry out the study as described, with 60 participants. What would be the power of the researcher's test to detect a mean difference in blood pressure of just 2 mm Hg?

Exercise 9 *Sample size*

Suppose that the psychologist in Activity 20 was planning a similar study on a different specific population, and that using a 5% significance level, the psychologist wished to detect a difference in mean IQ of 5 points between the specific population and the UK general population, with power 0.8. How many participants should the psychologist test?

Summary

This unit has considered the problem of testing hypotheses. The main idea is to consider two competing hypotheses: the null hypothesis and the alternative hypothesis. Data are observed relevant to the hypotheses, and summarised via the test statistic, whose distribution is known as the null distribution when the null hypothesis is true. The null distribution is then used to assess how likely the observed data would be if the null hypothesis were true.

Two approaches to testing hypotheses were considered. The first approach fixed the significance level which was then used to define a rejection region such that the null hypothesis is rejected if the test statistic is observed in the rejection region. The second approach used p-values, which give the probability of observing a result 'at least as extreme as' the result that was observed. The smaller the p-value, the more evidence there is against the null hypothesis.

Some standard tests were discussed: (one-sample) z- and t-tests, and a large sample test for the value of a proportion. These tests were carried out both by hand and using Minitab. Links between

- confidence intervals and hypothesis testing
- p-values and rejection decisions

were discussed, the latter in detail.

The final section of the unit introduced the concepts of Type I and Type II errors, together with the power of a test. A Type I error is where the null hypothesis is rejected when it is true, while a Type II error is when the null hypothesis is false, but it is not rejected. The power of a test is the

probability of rejecting the null hypothesis when it is false; as such, a high power is desirable.

One way to achieve high power without increasing the probability of a Type I error is by increasing the sample size. Calculating the power in a particular testing situation was considered, together with a method for choosing the best sample size in this situation. Minitab was also used for calculating the power and choosing the sample size, including for a less restrictive situation.

Learning outcomes

After you have worked through this unit, you should be able to:

- specify appropriate null and alternative hypotheses
- specify an appropriate null distribution for a test statistic
- calculate the rejection region for a test when the null distribution follows a normal or t-distribution
- appreciate the link between hypothesis tests and confidence intervals
- calculate a p-value from statistical tables when the null distribution follows a normal distribution
- use either the rejection region or the p-value to arrive at the correct conclusion for a test, and be able to communicate the conclusion in non-technical language
- understand how p-values can be related to tests at fixed significance levels
- appreciate common tests for:
  - testing the mean of a population (namely, the z-test and the t-test)
  - testing a proportion with a large sample
- understand the concepts of Type I and Type II errors
- understand the concept of the power of a test and the power function
- calculate the power of a test in the particular case in which a sample is drawn from a normally distributed population with known standard deviation
- calculate the sample size required to achieve a specified power for a test in the particular case in which a sample is drawn from a normally distributed population with known standard deviation
- use Minitab to:
  - carry out some standard hypothesis tests
  - make power calculations
  - make sample size calculations.

Solutions to activities

Solution to Activity 1

Let μ denote the mean pass rate nationally over the period April 2014–March 2015. The national pass rate for the period April 2013–March 2014 was 51.6%, so in testing the claim that μ is lower than the national pass rate for the same period the previous year, the null hypothesis would be that μ is no different to the national pass rate for the previous year, so the null hypothesis is $H_0 : \mu = 51.6\%$.

Solution to Activity 2

(a) Let μ_S be the mean systolic BP for women taking blueberry powder. The mean systolic BP for all post-menopausal women is 138 mm Hg, so the null hypothesis for testing whether the mean systolic BP for the blueberry group has decreased is $H_0 : \mu_S = 138$ mm Hg.

(b) Let μ_D be the mean diastolic BP for women taking blueberry powder. The mean diastolic BP for all post-menopausal women is 80 mm Hg, so the null hypothesis for testing whether the mean diastolic BP for the blueberry group has decreased is $H_0 : \mu_D = 80$ mm Hg.

Solution to Activity 3

For this test, we would like to detect whether μ_D is less than 80 mm Hg, so a suitable alternative hypothesis would be $H_1 : \mu_D < 80\,\text{mm Hg}$.

Solution to Activity 4

Because we want to be able to detect whether μ is lower than 51.6%, the alternative hypothesis would be $H_1 : \mu < 51.6\%$.

Solution to Activity 5

Because we want to be able to detect whether p is different to 0.25, the alternative hypothesis would be $H_1 : p \neq 0.25$.

Solution to Activity 6

(a) The null and alternative hypotheses for the test were

$$H_0 : \mu = 2.3, \quad H_1 : \mu \neq 2.3.$$

(b) The sample data used for the test were the sample mean $\overline{x} = 0.37$ and the sample standard deviation $s = 1.52$. The test statistic was the sample mean $\overline{X}$.

(c) The null distribution of the test statistic is the distribution of the test statistic when H_0 is true. This was

$$\overline{X} \approx N(2.3, 0.0118).$$

(d) The significance level of the test was 5%.

(e) The critical values of the test were $c_1 = 2.087$ and $c_2 = 2.513$. The rejection region for the test was then defined to be all values of $\overline{x}$ which are less than or equal to 2.087, or greater than or equal to 2.513. The rejection region is the set of values of the test statistic for which we will reject H_0.

(f) Because 0.37 is in the rejection region, this leads us to reject H_0 at the 5% significance level.

(g) We conclude that the data suggest that the mean festive weight gain is not equal to 2.3 kg, and the fact that the observed sample mean is so much lower than 2.3 kg would suggest that the mean festive weight gain is in fact lower than 2.3 kg.

Solution to Activity 7

(a) We wish to test whether the mean festive weight gain is 0.55 kg, so the null and alternative hypotheses are

$$H_0 : \mu = 0.55, \quad H_1 : \mu \neq 0.55.$$

(b) If H_0 is true, the weight gains will have a mean of 0.55 with some variance σ^2, so by the Central Limit Theorem when there are $n = 195$ observations,

$$\overline{X} \approx N\left(0.55, \frac{\sigma^2}{195}\right).$$

As before, s^2 can be used as an estimate for σ^2, so if H_0 is true, then

$$\overline{X} \approx N(0.55, 0.0118).$$

(c) It follows from part (b) that

$$Z = \frac{\overline{X} - 0.55}{\sqrt{0.0118}} \approx N(0,1),$$

so Z can be used as the test statistic. The observed value of Z is then

$$z = \frac{\overline{x} - 0.55}{\sqrt{0.0118}} = \frac{0.37 - 0.55}{\sqrt{0.0118}} \simeq -1.657.$$

(d) The test statistic Z is approximately distributed as $N(0,1)$ if H_0 is true, so the null distribution is $N(0,1)$. Using a 5% significance level, the critical values are then

$$c_1 = q_{0.025} = -1.960, \quad c_2 = q_{0.975} = 1.960.$$

Thus the rejection region is all values of z less than or equal to -1.960 or greater than or equal to 1.960.

(e) Since the observed value of the test statistic is $z = -1.657$, and $-1.960 < -1.657 < 1.960$, there is no evidence to reject H_0 at the 5% significance level.

(f) We conclude that there is no evidence to suggest that the mean festive weight gain is not 0.55 kg.

Solution to Activity 8

(a) We wish to test whether the mean festive weight gain is lower than 0.55 kg, and so the null and alternative hypotheses are

$$H_0 : \mu = 0.55, \quad H_1 : \mu < 0.55.$$

(b) The null hypothesis is the same as in Activity 7, and so, if H_0 is true, the test statistic is

$$Z = \frac{\overline{X} - 0.55}{\sqrt{0.0118}} \approx N(0, 1).$$

The observed value of z is also the same as in Activity 7, namely $z = -1.657$.

(c) This time there is a single critical value c_1 such that

$$P(Z \leq c_1) = 0.05,$$

so that

$$c_1 = q_{0.05} = -q_{0.95} = -1.645.$$

The rejection region is therefore all values of z less than or equal to -1.645.

(d) Since the observed value of the test statistic is $z = -1.657$ and $-1.657 < -1.645$ (just), the observed value of the test statistic is in the rejection region. We should therefore reject H_0 at the 5% significance level.

Note that in this one-sided version of the test in Activity 7, there is a different rejection region which has led to a different conclusion to the test, even though we have the same data.

(e) We conclude that there is evidence to suggest that the mean festive weight gain is lower than 0.55 kg.

Solution to Activity 9

(a) (i) The test statistic is

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{54.166 - 54.7}{2.864/\sqrt{137}} \simeq -2.182.$$

(ii) For a 5% significance level, the critical values are

$$c_1 = q_{0.025} = -q_{0.975} = -1.960 \quad \text{and} \quad c_2 = q_{0.975} = 1.960.$$

Hence the rejection region is values of z such that $z \leq -1.960$ or $z \geq 1.960$.

(iii) Since $-2.182 < -1.960$, the observed value of the test statistic is in the rejection region, so we reject H_0 at the 5% significance level. We conclude that the data suggest that the average driving theory test pass rate for females nationally over the period April 2014–March 2015 is different to the national pass rate for females for the same period the previous year. What's more, because the observed value of the test statistic is in the lower tail, reflecting the fact that the sample mean, 54.166, is smaller than the

hypothesised mean, 54.7, it looks like the average pass rate nationally for females is lower than the national female pass rate for the previous year.

(b) (i) The test statistic is

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{48.683 - 48.8}{2.336/\sqrt{137}} \simeq -0.586.$$

(ii) Here we have a one-sided test, so for a 10% significance level the critical value is

$$c_1 = q_{0.1} = -q_{0.9} = -1.282.$$

Hence the rejection region is values of z such that $z \leq -1.282$.

(iii) Since $-1.282 < -0.586$, the observed value of the test statistic is not in the rejection region, so we do not reject H_0 at the 10% significance level. We conclude that the data do not suggest that the average driving theory test pass rate for males nationally over the period April 2014–March 2015 is lower than the national pass rate for males for the same period the previous year.

(c) The results from parts (a) and (b) suggest that the lower mean pass rate nationally observed in Example 5 seems to be due to a decrease in the mean pass rate for females and not for males.

Solution to Activity 10

(a) The test statistic is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{75 - 80}{9/\sqrt{20}} \simeq -2.485.$$

(b) Here we have a one-sided test, so for a 5% significance level the critical value c_1 is the 0.05-quantile of the $t(20 - 1) = t(19)$ distribution, so that $c_1 = -1.729$ and the rejection region is all values of t such that $t \leq -1.729$.

(c) Since $-2.485 < -1.729$, the observed value of t is in the rejection region, so we reject H_0 at the 5% significance level. We conclude that the data suggest that taking 22 g of blueberry powder each day for 8 weeks lowers diastolic BP in menopausal women.

Solution to Activity 11

The sample size is large, so Z_p is an appropriate test statistic with null distribution $N(0, 1)$. Then

$$z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{111}{307} - 0.25}{\sqrt{\frac{0.25 \times 0.75}{307}}} \simeq 4.514.$$

The test is two-sided, so using a 5% significance level, the critical values are

$$c_1 = -1.960, \quad c_2 = 1.960,$$

and the rejection region is values of z_p such that $z_p \leq -1.960$ or $z_p \geq 1.960$.

Since $1.960 < 4.514$, the observed value of z_p is in the rejection region, so we reject H_0 at the 5% significance level. The data therefore suggest that the proportion of young adults aged 20–34 in Northern Ireland living with their parents in 2013 was not equal to 0.25. What's more, the fact that the observed value of the test statistic is in the upper tail suggests that the proportion of young adults living with their parents in Northern Ireland in 2013 was in fact greater than 0.25.

Solution to Activity 12

Since the alternative hypothesis is $H_1: \mu_M < 48.8$, values in the lower tail only will provide evidence against H_0. So any value of z such that $z \leq -0.586$ will be considered 'at least as extreme as' the observed value of the test statistic. Then

$$p = P(Z \leq -0.586),$$

where $Z \sim N(0,1)$. This can be calculated from the table of standard normal probabilities after appropriate rounding:

$$\begin{aligned} p &= P(Z \leq -0.586) = P(Z \geq 0.586) \\ &= 1 - P(Z < 0.586) \simeq 1 - P(Z < 0.59) \\ &= 1 - \Phi(0.59) = 1 - 0.7224 = 0.2776. \end{aligned}$$

To aid understanding, the p-value is illustrated in Figure 20.

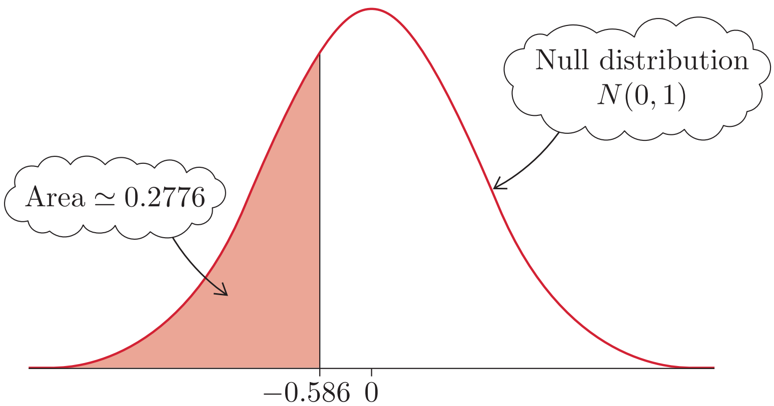


Figure 20 Null distribution $N(0,1)$ with the p-value for the observed value $z = -0.586$ marked

Since $p > 0.1$, there is little or no evidence against H_0. (Note that in Activity 9 the decision was made not to reject H_0 at the 10% significance level.)

Solution to Activity 13

The test is two-sided, so since the observed value of the test statistic was $z \simeq -1.657$, values of z such that $z \leq -1.657$ or $z \geq 1.657$ would be 'at least as extreme as' the value $z = -1.657$.

Thus the value of the p-value is

$$\begin{aligned} p &= P(Z \leq -1.657) + P(Z \geq 1.657) \\ &= P(Z \geq 1.657) + P(Z \geq 1.657) \\ &= 2P(Z \geq 1.657) \simeq 2P(Z \geq 1.66) \\ &= 2(1 - \Phi(1.66)) = 2(1 - 0.9515) = 2 \times 0.0485 = 0.097. \end{aligned}$$

The p-value is illustrated in Figure 21.

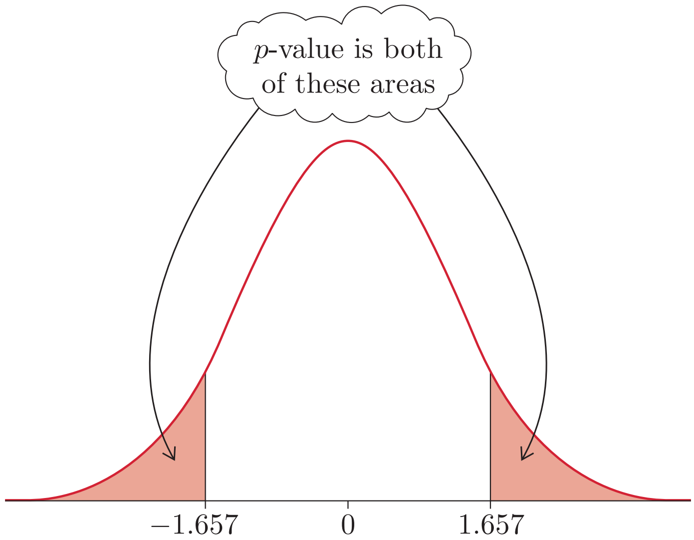


Figure 21 Null distribution $N(0, 1)$ with the two-sided p-value for the observed value $z = -1.657$ marked

The value of p is such that $0.05 < p < 0.10$, so from Table 3, there is weak evidence against H_0. (Note that in Activity 7 we did not reject H_0 at the 5% significance level.) Further, the fact that $z = -1.657$ suggests that there is weak evidence that the mean festive weight gain is lower than 0.55 kg.

Solution to Activity 14

If the test is two-sided, then

$$p = P(T \leq -1.841) + P(T \geq 1.841).$$

Since the $t(19)$ distribution is symmetric about 0,

$$P(T \leq -1.841) = P(T \geq 1.841),$$

so

$$p = 2P(T \leq -1.841).$$

But this is twice the p-value for the one-sided test in Example 12. Thus $p = 2 \times 0.041 = 0.082$.

Since $0.05 < p < 0.1$, there is now only weak evidence against the null hypothesis that mean systolic blood pressure is equal to 138 mm Hg. (In the one-sided version of the test in Example 12 there was moderate evidence against the null hypothesis.)

Solution to Activity 15

The p-value of $p = 0.082$ is such that $0.01 < 0.05 < p < 0.1$. It follows that H_0 would have been rejected by a test at the 10% significance level (because $p < 0.1$), and would not have been rejected by tests at either the 5% or 1% significance levels (because $p > 0.05$ and $p > 0.01$).

Solution to Activity 16

(a) The p-value is such that $0.01 < p \leq 0.05$. It follows that H_0 would have been rejected by tests at the 5% and 10% significance level (because $p \leq 0.05$ and $p < 0.1$), and would not have been rejected by a test at the 1% significance level (because $p > 0.01$).

(b) From Table 3, strong evidence against H_0 corresponds to $p \leq 0.01$. In this case, H_0 would have been rejected by tests at each of the 1%, 5% and 10% significance level (because $p \leq 0.01 < 0.05 < 0.1$).

Solution to Activity 17

Since the test results in 'reject H_0', regardless of the data, both the power (the probability of rejecting H_0 when H_0 is false) and the significance level (the probability of rejecting H_0 when H_0 is true) are 1. That's a great power! But it is achieved at an unacceptable cost. The test is unacceptable because the significance level, which should be small, is also 1 – this person would also always reject H_0 even when there is no evidence to do so.

Solution to Activity 18

Rearranging the relationship between W and Z_1 gives

$$Z_1 = W + \frac{d}{\sigma/\sqrt{n}}.$$

Also, from the text,

$$W \sim N(0, 1).$$

Then, setting $a = 1$, $X = W$ and $b = d/(\sigma/\sqrt{n})$ in the result quoted in the question,

$$Z_1 \sim N\left(1 \times 0 + \frac{d}{\sigma/\sqrt{n}}, 1^2 \times 1\right) = N\left(\frac{d}{\sigma/\sqrt{n}}, 1\right).$$

Solution to Activity 19

(a) Since $H_1 : \mu < \mu_0$, the critical value c is such that

$$P(Z_1 \leq c \text{ when } H_0 \text{ true}) = \alpha.$$

But if H_0 is true, then $Z_1 \sim N(0, 1)$, so c is $-q_{1-\alpha}$ and the rejection region is all values of z_1 such that $z_1 \leq -q_{1-\alpha}$.

(b) If $\mu = \mu_0 - d$, $d > 0$, then
$$\overline{X} \sim N\left(\mu_0 - d, \frac{\sigma^2}{n}\right),$$
so
$$\frac{\overline{X} - (\mu_0 - d)}{\sigma/\sqrt{n}} \sim N(0,1).$$
But
$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} + \frac{d}{\sigma/\sqrt{n}} = Z_1 + \frac{d}{\sigma/\sqrt{n}} = V \sim N(0,1).$$

(c) We have
$$\begin{aligned}
\text{power} &= P(Z_1 \leq -q_{1-\alpha} \text{ when } \mu = \mu_0 - d) \\
&= P\left(Z_1 + \frac{d}{\sigma/\sqrt{n}} \leq -q_{1-\alpha} + \frac{d}{\sigma/\sqrt{n}}\right) \\
&= P\left(V \leq -\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right)\right) \\
&= \Phi\left(-\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right)\right) \\
&= 1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right)
\end{aligned}$$
since $V \sim N(0,1)$ and, as usual, $\Phi(-x) = 1 - \Phi(x)$ for any x.

Solution to Activity 20

The probability that the psychologist will reject the null hypothesis is the value of the power when the true value of μ is 104.75.

For this test, $\alpha = 0.05$ and the test is two-sided, so $q_{1-(\alpha/2)} = q_{0.975} = 1.960$. The sample size is $n = 80$ and the population standard deviation is $\sigma = 15$. The actual mean is assumed to be $\mu = \mu_0 + d = 104.75$, where $\mu_0 = 100$ (the mean of the scores in the general UK population). Thus $d = 104.75 - 100 = 4.75$. Therefore
$$\begin{aligned}
\text{power} &= 1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(1.960 - \frac{4.75}{15/\sqrt{80}}\right) \\
&= 1 - \Phi(-0.872) \simeq 1 - \Phi(-0.87) = \Phi(0.87) = 0.8078.
\end{aligned}$$

The psychologist's test therefore has a reasonably good chance of finding a difference in mean IQ of this size.

Solution to Activity 21

When the true mean equals the hypothesised mean, the null hypothesis is true. And the probability of rejecting H_0 when the null hypothesis is true is the significance level 0.05.

Solution to Activity 22

The power function for the two-sided test is 'U-shaped', increasing as the true value of μ becomes both larger and smaller than the hypothesised value $\mu_0 = 5$. This is because for a two-sided test we are interested in differences from μ_0 in either direction.

Notice that the power function is still minimised when the true mean equals the hypothesised mean, that is, when the null hypothesis is true. As with the one-sided test, in this case the power is 0.05, the significance level.

Solution to Activity 23

(a) According to Equation (4), the sample size is proportional to the variance σ^2 and hence should increase if the amount of variability in the data increases (as seems intuitively reasonable). More specifically, if the standard deviation σ is changed to 2σ, then n should be multiplied by $2^2 = 4$.

(b) According to Equation (4), the sample size is inversely proportional to the square of d, the difference between means, and hence should increase if the difference between means that you wish to detect is decreased (this also seems intuitively reasonable because you are trying to identify something less clear-cut). More specifically, if d is changed to $d/2$, then n should be divided by $(1/2)^2 = 1/4$, that is, multiplied by 4.

Solution to Activity 24

(a) From Activity 19, the expression for the power when the alternative hypothesis is $H_1 : \mu < \mu_0$ is the same as the expression for the power when the alternative hypothesis is $H_1 : \mu > \mu_0$. Therefore tests for either alternative hypothesis will be rearranged to give the same expression for n when setting the power to be the value γ.

(b) From Equation (2), the only difference between the expression for the power when the alternative hypothesis is $H_1 : \mu \neq \mu_0$ and when the alternative hypothesis is $H_1 : \mu < \mu_0$ (or $H_1 : \mu > \mu_0$) is that $q_{1-(\alpha/2)}$ appears in the expression rather than $q_{1-\alpha}$. The expression for n will therefore be the same, except that $q_{1-\alpha}$ will be replaced by $q_{1-(\alpha/2)}$.

Solution to Activity 25

This is a two-sided test, so the appropriate number of participants is

$$n = \frac{\sigma^2}{d^2}\left(q_{1-(\alpha/2)} - q_{1-\gamma}\right)^2.$$

Here $\alpha = 0.01$. The assumed population standard deviation σ is 10. The difference d that the test is being designed to detect is 5, and γ, the required power to detect such a difference, is 0.9. Thus $q_{1-(\alpha/2)} = q_{0.995}$ $= 2.576$ and $q_{1-\gamma} = q_{0.1} = -q_{0.9} = -1.282$. So the sample size required is

$$n = \frac{10^2}{5^2}(2.576 + 1.282)^2 \simeq 59.5,$$

which is rounded up to 60.

Solutions to exercises

Solution to Exercise 1

(a) Let μ_F denote the mean pass rate for females nationally over the period April 2014–March 2015. The national pass rate for females for the period April 2013–March 2014 was 54.7%, so in testing the claim that μ_F is different to the national pass rate for females for the same period the previous year, the null hypothesis would be that μ_F is no different to the national pass rate for females for the previous year, so the null hypothesis is $H_0 : \mu_F = 54.7\%$.

For this test we would like to detect any difference for μ_F both greater than or less than 54.7%, so a suitable alternative hypothesis would be $H_1 : \mu_F \neq 54.7\%$.

(b) Let μ_M denote the mean pass rate nationally for males over the period April 2014–March 2015. The national pass rate for males for the period April 2013–March 2014 was 48.8%, so in testing the claim that μ_M is lower than the national pass rate for males for the same period the previous year, the null hypothesis would be that μ_M is no different to the national pass rate for males for the previous year, so the null hypothesis is $H_0 : \mu_M = 48.8\%$.

For this test we would like to detect whether μ_M is lower than 48.8%, so a suitable alternative hypothesis would be $H_1 : \mu_M < 48.8\%$.

Solution to Exercise 2

The first step in carrying out a z-test is to specify the null and alternative hypotheses. We wish to test whether μ is greater than 1, so the hypotheses are

$$H_0 : \mu = 1, \quad H_1 : \mu > 1.$$

The test statistic for a z-test is

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.636 - 1}{1.655/\sqrt{33}} \simeq 2.208.$$

This is a one-sided test and therefore the critical value for a 5% significance level is $c_2 = 1.645$, so the rejection region is all values of z such that $z \geq 1.645$.

Since $1.645 < 2.208$, the observed value of z lies in the rejection region and we reject H_0 at the 5% significance level. We conclude that the data suggest that the mean number of insects of the taxon *Staphylinoidea* in the traps is greater than 1.

Solution to Exercise 3

We wish to test whether the population mean spelling test score, μ, of visually impaired children using braille is 100, and we are interested in detecting departures from 100 in either direction. Thus the hypotheses are

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100.$$

A normal model for the data has been assumed, so the t-test would be an appropriate test to use. The test statistic for a t-test is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{99.0 - 100}{11.7/\sqrt{23}} \simeq -0.410.$$

This is a two-sided test and therefore the critical values are the 0.025-quantile and the 0.975-quantile of the $t(23 - 1) = t(22)$ distribution. So, from the table in the Handbook, $c_1 = -2.074$ and $c_2 = 2.074$, and hence the rejection region is all values of t such that $t \leq -2.074$ or $t \geq 2.074$.

Since $-2.074 < -0.410 < 2.074$, the observed value of the test statistic is not in the rejection region so we do not reject H_0 at the 5% significance level. We conclude that there is no evidence to suggest that the mean spelling score of visually impaired children using braille is not 100, the mean value in the sighted population.

Solution to Exercise 4

The sample size is large, so Z_p is an appropriate test statistic with null distribution $N(0, 1)$. Then $\widehat{p} = x/n$ and

$$z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{86}{344} - 0.25}{\sqrt{\frac{0.25\times 0.75}{344}}} = 0.$$

Zero has appeared because the proportion of young adults aged 20–34 years in Scotland living with their parents in the sample equates precisely to the proportion of young adults aged 20–34 years in the UK living with their parents: $86/344 = 0.25$. The test is two-sided, so using a 5% significance level, the critical values (from the table of normal quantiles) are

$$c_1 = -1.960, \quad c_2 = 1.960,$$

and the rejection region is values of z_p such that $z_p \leq -1.960$ or $z_p \geq 1.960$.

Since $-1.960 < 0 < 1.960$, the observed value of z_p is not in the rejection region, so we do not reject H_0 at the 5% significance level. In fact, the observed value of the test statistic is directly in the centre of the null distribution and as far from the tails as it possibly can be! The data therefore provide no evidence to suggest that the proportion of young adults living with their parents in Scotland in 2013 was not 0.25, unsurprisingly, given that the data give rise to the same proportion as that being tested.

Solution to Exercise 5

This is a one-sided test, so all values of z such that $z \geq 2.208$ will be at least as extreme as the observed value of the test statistic. So

$$\begin{aligned} p &= P(Z \geq 2.208) = 1 - P(Z < 2.208) \\ &= 1 - \Phi(2.208) \simeq 1 - \Phi(2.21) = 1 - 0.9864 = 0.0136. \end{aligned}$$

Since $0.01 < p < 0.05$, there is moderate evidence against H_0, so we conclude that there is moderate evidence that the mean number of insects of the taxon *Staphylinoidea* is greater than 1.

Solution to Exercise 6

(a) This is a two-sided test, so all values of t such that $t \leq -0.410$ and $t \geq 0.410$ will be at least as extreme as the observed value of the test statistic. So

$$p = P(T \leq -0.410) + P(T \geq 0.410) = 2P(T \geq 0.410),$$

where T follows the t-distribution with $n - 1 = 23 - 1 = 22$ degrees of freedom.

(b) Since $p = 0.6858 > 0.10$, there is little or no evidence against H_0, so we conclude that there is little or no evidence that the mean spelling score of visually impaired children using braille is not 100.

Solution to Exercise 7

The p-value is such that $0.05 < p \leq 0.10$. It follows that H_0 would have been rejected by a test at the 10% significance level (because $p \leq 0.1$), and would not have been rejected by tests at the 1% and 5% significance levels (because $p > 0.01$ and $p > 0.05$).

Solution to Exercise 8

Here, in the notation of power calculations, we have

$$\alpha = 0.01, \quad d = 2, \quad \sigma = 10, \quad n = 60.$$

The test is two-sided, so the required power is given by Equation (2):

$$\begin{aligned} \text{power} &= 1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(q_{0.995} - \frac{2}{10/\sqrt{60}}\right) \\ &= 1 - \Phi\left(2.576 - 1.549\right) \simeq 1 - \Phi(1.03) \\ &= 1 - 0.8485 = 0.1515 \simeq 0.152. \end{aligned}$$

This value is low, so the procedure does not have a very good chance of detecting a difference in blood pressure of 2 mm Hg.

Solution to Exercise 9

Here, in the notation of sample size calculations, we have

$$\alpha = 0.05, \quad d = 5, \quad \sigma = 15, \quad \gamma = 0.8.$$

This is a two-sided test, so $q_{1-(\alpha/2)} = q_{0.975} = 1.960$ and $q_{1-\gamma} = q_{0.2} = -q_{0.8} = -0.8416$. Then the sample size required is

$$\begin{aligned} n &= \frac{\sigma^2}{d^2}\left(q_{1-(\alpha/2)} - q_{1-\gamma}\right)^2 \\ &= \frac{15^2}{5^2}(1.960 + 0.8416)^2 \simeq 70.6, \end{aligned}$$

which is rounded up to 71.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 204: © iStockphoto.com/mrPliskin

Page 205: Hongqi Zhang/123rf.com

Page 207: https://www.gov.uk/government/news/review-of-foreign-languages-on-driving-tests Reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence

Page 208: © nickyp2/www.istockphoto.com

Page 213: Minitab Inc.

Page 214: © Francois Nel/Getty Images.com

Page 216: Pablo Andres. This file is licensed under the Creative Commons Attribution-Non-commercial-No Derivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0

Page 218: Taken from: http://sleepcenterlcmc.com/category/fitness-and-health

Page 220: Taken from: www.nepalikisan.com

Page 221: United States Department of Agriculture, Agricultural Research Service

Page 223: © iStockphoto.com/PeopleImages

Page 225 top: Sarefo. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 225 bottom: Taken from: http://www.wikiwand.com/en/English_Braille

Page 227: InterestingPics. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 229 top: Sergey Furtaev/www.123rf.com

Page 229 bottom: szefei/www.123rf.com

Page 231: © 2017 FreezeDry

Page 233: Shabib Khan. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by-nc/2.0/

Page 235: © sturti/iStock/Getty Images Plus

Page 238: sportgraphic/www.123rf.com

Page 240: evrenkalinbacak/www.123rf.com

Page 241: Tony Hiscott. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 244: Adrian Pingstone

Page 248: Department of Psychology, Colombia University

Page 254: auremar/www.123rf.com